

CENTRAL LIBRARY

**Birla Institute of Technology & Science
PILANI (Rajasthan)**

Call No.

Accession No.

Acc. No.....

ISSUE LABEL

Not later than the latest date stamped below.

--	--	--

PRENTICE-HALL PSYCHOLOGY SERIES

F. A. MOSS, Ph.D., M.D., EDITOR

Measurement
in
Psychology

PRENTICE-HALL PSYCHOLOGY SERIES

F. A. Moss, Ph.D., M.D., EDITOR

GENERAL PSYCHOLOGY, *by* Floyd C. Dockeray, Ph.D.

STUDENT'S GUIDE FOR BEGINNING THE STUDY OF PSYCHOLOGY, *by* Valentine, Taylor, Baker, *and* Stanton.

FOUNDATIONS OF ABNORMAL PSYCHOLOGY, *by* Fred A. Moss, Ph.D., M.D., *and* Thelma Hunt, Ph.D., M.D.

A WORK BOOK IN EDUCATIONAL PSYCHOLOGY, *by* Harvey C. Lehman, Ph.D., *and* Stuart M. Stoke, Ed.D.

CHILD PSYCHOLOGY, *by* Arthur T. Jersild, Ph.D.

THE PSYCHOLOGY OF ADOLESCENCE, *by* Karl C. Garrison, Ph.D.

SOCIAL PSYCHOLOGY, *by* Abraham Myerson, M.D.

THE PRINCIPLES AND METHODS OF VOCATIONAL CHOICE, *by* Maurice J. Neuberg, Ph.D.

COMPARATIVE PSYCHOLOGY, *by* Thorndike, Waters, Stone, Moss, Purdy, Fields, Franz, Liddell, Heron, Tolman, Tryon, *and* Tinklepaugh.

STUDY OUTLINE FOR GENERAL PSYCHOLOGY, *by* S. L. Crawley, Ph.D.

STUDYING EFFICIENTLY, *by* S. L. Crawley, Ph.D.

PERSONALITY, *by* Harold V. Gaskill, Ph.D.

LEGAL PSYCHOLOGY, *by* Harold E. Burt, Ph.D.

INTRODUCTION TO EXPERIMENTAL PSYCHOLOGY, *by* Paul F. Finner, Ph.D.

EDUCATIONAL PSYCHOLOGY, *by* Skinner, McConnell, Lawther, Hartmann, Gray, Fletcher, Thomson, Jersild, Powers, Boynton, Conklin, Davis, Webb, Garrison (K.C.), Freeman, Witty, Lincoln, Wood, Gifford, Rock, Wallin, Moss, Trabue, Aleck, *and* Washburne.

FUNDAMENTALS OF PSYCHOLOGY IN SECONDARY EDUCATION, *by* S. C. *and* K. C. Garrison.

MEASUREMENT IN PSYCHOLOGY, *by* Thelma Hunt, Ph.D., M.D.

EDUCATIONAL PSYCHOLOGY, *by* Thomas R. Garth, Ph.D.

THE TWO SCIENCES OF PSYCHOLOGY, *by* Arthur D. Fearon, Ph.D.

Measurement
in
Psychology

by

Thelma Hunt, Ph.D., M.D.

*Assistant Professor of Psychology
The George Washington University*

NEW YORK
PRENTICE-HALL, INC.
1937

COPYRIGHT, 1936, BY
PRENTICE-HALL, INC.

ALL RIGHTS RESERVED. NO PART OF THIS BOOK MAY
BE REPRODUCED IN ANY FORM, BY MIMEOGRAPH OR ANY
OTHER MEANS, WITHOUT PERMISSION IN WRITING FROM
THE PUBLISHERS.

First PrintingOctober, 1936
Second PrintingJune, 1937

PRINTED IN THE UNITED STATES OF AMERICA

TO MY VARIOUS TEACHERS AND INSTRUCTORS

WHO HAVE BEEN RESPONSIBLE FOR THE BACKGROUND
OF TRAINING AND EXPERIENCE WHICH MAKE
POSSIBLE THE WRITING OF THIS BOOK

Preface

THIS book is designed primarily as a textbook for college courses in psychological tests and measurements. The aim has been to give a brief survey of the whole field of psychological testing. In this respect the book will be found to differ from the several books which treat only one phase of psychological measurement, as the intelligence, the educational, or some other type of testing.

A basic theme of the book is that quantitative study and measurement are just as pertinent to psychological pursuits as to other scientific pursuits; and that progress in psychology will generally be in proportion to the use of objective quantitative methods of study.

It is difficult to give credit to all those who have assisted in the preparation of this book. I wish to express my appreciation to the writers and publishers who have permitted me to quote from their works. Specific acknowledgment is made to them at appropriate places in the text. I am indebted to my students in tests and measurements, not only for assistance in collecting facts and information, but also for their patience and interest in connection with my try-out, in teaching, of most of the material of the book. Dr. Fred A. Moss has contributed much through his guidance and encouragement during the writing of the book. I am under particular obligation to Dr. Katharine Omwake, of Agnes Scott College, who read the whole manuscript and made many valuable suggestions; and to Dr. L. J. O'Rourke, of the United States Civil Service Commission, who furnished suggestions and assistance in the preparation of the

chapters dealing with tests in the field of personnel work. Finally, I wish to thank Miss Margaret Telford and Mrs. Sarah Menzer for assistance in the preparation of the manuscript.

THELMA HUNT

The George Washington University

Contents

PART I

THE PLACE OF MEASUREMENT IN PSYCHOLOGY

	PAGE
PREFACE	vii
CHAPTER	
I. THE VALUE OF MEASUREMENT IN PSYCHOLOGY	3
What is measurement?	4
Measurement in everyday life	5
Origin and development of measurement	6
Value of measurement	7
Problems of measurement in psychology	11
II. THE INSTRUMENTS OF MEASUREMENT IN PSYCHOLOGY	13
Psychological tests	14
Rating scales	16
Questionnaires	17
Raw measures and relative measures	18
Reliability and validity	19
III. THE HISTORY OF MEASUREMENT IN PSYCHOLOGY	23
The background of intelligence test development—the pre-Binet period	25
The Binet period	28
Preliminary period of group mental tests	32
The World War period—the Army tests	34
The post-War period—application of Army tests and development of similar tests	38
The maturing and stabilizing period of 1921-1926	34
The present period	41

PART II

MEASUREMENT OF INTELLECTUAL QUALITIES

CHAPTER	PAGE
IV. THE MEASUREMENT OF MENTAL DEFICIENCY	59
What is mental deficiency?	59
Degrees and types of mental deficiency . .	61
Why measure mental deficiency?	62
Types of tests suitable for measuring mental deficiency	63
V. THE MEASUREMENT OF SUPERIORITY	82
Development of interest in the superior .	85
The superior school child	91
Tests suitable for testing school groups .	95
Mental tests for the college level	104
Measuring the superiority of adults outside of college	107
VI. INTELLECTUAL MEASUREMENT OF THE INSANE	109
General mental tests	110
Memory tests	114
Association tests	115
VII. THE USES OF MENTAL TESTS	118
Uses in college	119
Uses in schools below the college level .	123
Uses of mental tests in vocational selection	124
Uses of mental tests with delinquents and criminals	133
The use of mental tests in various theoretical and research problems	136

PART III

MEASUREMENT OF APTITUDES

VIII. THE MEASUREMENT OF SPECIAL TALENTS . .	143
The measurement of musical aptitude . .	144
Measurement in the field of art	153

CHAPTER	PAGE
IX. THE MEASUREMENT OF MECHANICAL APTITUDE	162
The Stenquist Mechanical Aptitude Tests	164
The Minnesota Mechanical Ability Tests	167
The MacQuarrie Test for Mechanical Ability	169
The O'Rourke Mechanical Aptitude Test	171
X. THE MEASUREMENT OF APTITUDE FOR VOCATIONS AND PROFESSIONS	172
The problem	174
The development of the aptitude test	174
The nature of the aptitude test	175
The validity of the test	180
XI. THE MEASUREMENT OF INTERESTS	187
The methods of measuring interests	189

PART IV

MEASUREMENT OF ACHIEVEMENT

XII. THE MEASUREMENT OF ACHIEVEMENT IN SCHOOLS	207
What are tests of achievement?	207
Psychology's contribution to achievement testing	208
Examples of achievement tests for schools	213
The nature of short-answer achievement tests	223
XIII. OBJECTIVE ACHIEVEMENT TESTS IN PROFESSIONAL SCHOOLS	227
The difficulties encountered	227
Early studies in achievement testing in professional schools	228
An objective bacteriology test	229
Study and standardization of the tests in bacteriology	236

CHAPTER	PAGE
XIV. THE MEASUREMENT OF JOB EFFICIENCIES . .	240
A service rating scale	241
Objective psychological tests in the meas- urement of job efficiencies	250

PART V

MEASUREMENT IN INDUSTRIAL AND
PERSONNEL FIELDS

XV. HISTORICAL BACKGROUND OF PSYCHOLOGICAL MEASUREMENT IN INDUSTRY	257
The foundations of psychology applied to personnel problems	257
Early experiments in industrial psychology	262
The effect of the World War	265
History of the testing of applicants . . .	266
XVI. CONSTRUCTING A "PSYCHOLOGICAL" TEST FOR EMPLOYMENT	273
Analysis of the job	275
Selection of types of tests and questions .	276
Length of test	277
Construction of questions	277
Analysis of test material from preliminary trial	278
Final arrangement of test	279
Final application of test and establishment of critical scores	279
XVII. MEASUREMENT IN THE SELECTION OF EM- PLOYEES	280
The purpose of measurement in selection .	280
Problems related to measurement in the se- lection of employees	281
Measurement in the selection of employees illustrated	286

Contents

xiii

CHAPTER	PAGE
XVIII. PSYCHOLOGICAL MEASUREMENT IN THE CONTROL OF EMPLOYEES	303
Example of measurement in relation to accident prevention	306
Time and motion study	309
Measurement in department store management	311

PART VI

MEASUREMENT OF THE MORE GENERAL TRAITS OF PERSONALITY

XIX. THE NATURE OF MEASUREMENT OF PERSONALITY	321
The difficulties of measuring personality traits	322
Methods of personality testing	325
XX. THE MEASUREMENT OF SOCIAL ATTRIBUTES	334
The Moss-Hunt-Omwake Social Intelligence Test	335
The Vineland Social Maturity Scale	351
XXI. THE MEASUREMENT OF EMOTIONS BY VERBAL TESTS	357
Adjustment questionnaires and tests	358
Extroversion-introversion scales	362
An annoyance test	365
The Presscy X-O test	367
XXII. THE MEASUREMENT OF CHARACTER AND THE MORAL SENSE	372
The tests of conduct knowledge and judgment	375
Behavior tests	380
Other types of tests	386

PART VII

PHYSIOLOGICAL MEASUREMENTS IN
PSYCHOLOGY

CHAPTER	PAGE
XXIII. THE MEASUREMENT OF FATIGUE	393
Blood pressure	398
Carbon-dioxide combining power of the blood	399
Blood sugar	399
Metabolism	400
Blood cell studies	400
XXIV. LABORATORY TESTS IN MENTAL DISORDERS .	403
Blood tests	404
Spinal fluid tests	406
Basal metabolism	408
Blood pressure	409
XXV. GLANDULAR FUNCTION TESTS	410
Nature of the endocrine glands and their importance to psychology	410
Measurement in relation to endocrine func- tion	416
XXVI. PHYSIOLOGICAL MEASUREMENTS OF EMOTION .	424
The nature of physiological measurements of emotions	425
Blood pressure measurements of emotion .	428
Measurement of breathing	431
Measurement of digestive-system activity .	432
Psychogalvanic reflex measurements . . .	433
Emotion indicated by measurement of ad- renalin in bloodstream	434
Physiological measurement of emotion in detection of lying	436
Summary	440

Contents

xv

CHAPTER	PAGE
XXVII. TESTS IN THE MOTOR AND SENSORY FIELDS .	441
Cephalic index	442
The strength tests	443
Tests for sensory acuity	446
Color blindness	448
Steadiness tests	448
Reaction time	452

List of Figures

	PAGE
1. Tests in the Pintner-Paterson Performance Scale	74
2. Test One—The International Intelligence Test	78
3. Porteus Mazes	79
4. The Distribution of Intelligence	92
5. Diagrammatic Representation of the Selective Process in Determination of Intelligence of School Groups	94
6. Mental Growth Curve	137
7. Samples from the Meier-Seashore Art Judgment Test	156
8. Samples from the McAdory Art Test	158-160
9. Sample from the Stenquist Mechanical Aptitude Test	167
10. Minnesota Paper Form Board	168
11. Sample Questions from the O'Rourke Mechanical Aptitude Test	170
12. Distribution of Medical School Grades According to Medical Aptitude Test Scores	181
13. Distribution of Intern Ratings According to Medical Aptitude Test Scores	183
14. Comparison of Various Criteria for Admission to Medical Schools	184
15. Distribution of Scores for Interest in Personnel Management	194
16. Samples from the Ayres Handwriting Scale	221
17. Distribution of Probst Service Ratings	247
18. Chart Showing Reliability of Probst Service Ratings	248
19. Sample Card from Munsterberg's Test for Street-Car Motormen	264
20. Relation Between Test Scores and Efficiency of Mail Distributors	293
21. Improved Selection Made by New Postal Examination	294

	PAGE
22. Circles Test for Dishonesty	382
23. The Relationship Between Production and Feelings of Tiredness	395
24. Blood Pressure Curves in Fatigue	398
25. Metabolism Records Before and After Fatigue	401
26. Physiological Changes Produced by Fatigue	402
27. Colloidal Gold Curves in Various Mental Disorders	407
28. Technique of Blood Estrin Determination	421
29. Blood Estrin Content at Various Stages of the Sex. Cycle	422
30. Diagram Showing Method Used to Record Stomach Contractions	427
31. Record Indicating Inhibitory Effect of Adrenalin on Intestinal Muscle Contraction	435
32. Head Calipers	442
33. Dynamometer for Measuring Hand Strength	444
34. Instrument for Testing Hand Steadiness	449
35. Wobblemeter for Measuring General Bodily Steadi- ness	450
36. Diurnal Variations in Steadiness	452
37. Moss-Brown Reaction Time Instrument	455

List of Tables

	PAGE
I. Army Test Ratings	37
II. Equivalent Mental Ages and I. Q.'s for Low Army Alpha Scores	81
III. Grade Norms (Average) for Scale A, Na- tional Intelligence Test	97
IV. Mental Age Equivalents for National Intel- ligence Test Scores	98
V. Norms for International Intelligence Test .	101
VI. Free Association Test (Kent-Rosanoff) . .	116
VII. Responses of Various Groups to Kent- Rosanoff Association Test	117
VIII. Uses Made of Mental Tests Among Staff Members at Purdue University	122
IX. Employment Methods Used by States . .	126
X. Intelligence Limits for Various Occupations	128
XI. Validity of Various Employment Methods .	131
XII. The Proportion of Successes and Failures at Each Mental Age Level for 413 Packing Jobs	132
XIII. Percentage of Various Criminals Who Are Mentally Inferior	135
XIV. Measures of the Variations in the I. Q. on Retesting as Found in Several Typical Studies	138
XV. Reliability Coefficients for Seashore Tests .	149
XVI. Validity Coefficients for Seashore Tests . .	152
XVII. Norms on the McAdory Art Test	161
XVIII. Correlations between Shop-Teacher Ranks and Series I Stenquist Tests	165
XIX. Correlations between Stenquist Assembly Tests and General Intelligence	166
XX. Per Cent Marks Assigned to a Physiology Examination	210
XXI. Records for Bacteriology Test Study . . .	238
XXII. Norms on Bacteriology Test	239

XXIII.	Comparison of Efficiencies of Instructors . .	252
XXIV.	Relationship of Social Intelligence Test Scores to Number of Extra-curricular Activities	347
XXV.	Relationship of Social Intelligence Test Scores to Activity Scores	348
XXVI.	Correlations between Social and Abstract Intelligence	349
XXVII.	Median Social Intelligence Test Scores for Various Occupational Groups	351
XXVIII.	Effects on Respiration of Fear Produced by Falling	431
XXIX.	Norms for Strength of Grip (in Kilograms)	445
XXX.	Correlations between Intelligence and Reaction Time	456

PART I

THE PLACE OF MEASUREMENT IN PSYCHOLOGY

CHAPTER I

The Value of Measurement in Psychology

WE ARE to concern ourselves throughout this text with measurement as it is applied to psychology. We are to examine the measuring tools which the psychologist employs. The aim of the text might be considered as twofold: to show the importance to psychology of exact measurement; and to familiarize the student with the types of psychological measurement that have been developed.

The text will cover more than intelligence and achievement tests; it will cover measurement in its widest variations. In regard to mental qualities, we shall consider not only the ordinarily discussed intelligence tests, but also the methods which have been developed for measuring special aptitudes and various traits of personality. In the field of achievement testing, we shall consider the tests for measuring achievement in business and industry as well as achievement in the schools. And finally, where they have been found particularly useful in psychological problems, we shall consider briefly tests borrowed from other sciences.

The method of discussion will consist in designating the several varieties of measurement in psychology, and in giving examples of the measuring devices which have been developed for them. In many instances, obviously, it will be impossible to mention all the means of measurement that have been developed; then we shall

4 Value of Measurement in Psychology

attempt to select those that are the more important or the more representative.

I. What is Measurement?

Things are measured when they are expressed quantitatively; or, we may say that a thing is measured when we speak of it in terms of "how much." To make clear what we mean by psychological measurement, let us examine a few instances.

Intelligence is measured when we can describe the amount of this quality which a person has. It is inexactly measured when we have means of indicating whether a person possesses genius intelligence, superior intelligence, average intelligence, subnormal intelligence, or intelligence at the feeble-minded level. It is more accurately measured when we have a measuring rod which enables us to designate intelligence by an Intelligence Quotient of 140, of 128, of 98, or some other definite quantitative amount.

To take another example, *achievement* is measured when we indicate by some measuring device just exactly the amount of achievement, or the degree of achievement, which a person has reached. This we may indicate by stating the number of problems he can solve or the number of correct judgments he can make in a given field. By other measuring devices the achievement may be stated in terms of how the individual compares in achievement with other individuals of similar age, of similar grade in school, or of similar occupation.

To mention one more example, *emotional reactions* are measured when we state the degree of an emotion that a person shows, or when we state how he compares in his emotional reactions with others. In the sphere of emotions our measurements are as yet very rough and

imperfect. In those of intelligence and achievement our methods have been refined to a fair degree of accuracy.

II. Measurement in Everyday Life

Since psychology observes man in all his activities and relationships, we may point out at the beginning the importance of measurement in man's everyday life. Measurement is always with us. There are no human activities that we can think of in which measurement does not enter in some way. At birth the child is measured for weight and other elements of physiological make-up that may indicate his possibilities for future health and life. At death a person is measured for length, so that he may be placed in a casket of proper size. After death, two dates are carved on his tombstone to indicate the measure of his life span.

In his daily activities man is continually guided by measurements. His whole day is planned by a device for measuring time. All his financial activities are carried on in terms of money, a measure of economic exchange. His travel is planned according to the measurements of distances. His illnesses are cared for by medicine given in measured doses. His bad eyesight is cured by a glass with a measured curve to fit a measured defect of his eye. Thus we are continually measuring our environment in order to make the most of it. Many more examples of measurement in man's everyday life might be put down, but these will serve to indicate its great importance.

The scope of measurement is almost without limit. Limits in size extend from the measurement of the largest planets in our solar system to the measurement of the smallest invisible constituents of matter. With ref-

6 Value of Measurement in Psychology

erence to nearness and remoteness, the measurements may be as near as the structures of our own body, or as remote as the farthest star in the heavens. In types of measuring devices, the "yardsticks" which are used, there is a variety too great to be discussed here. The type itself may be as definite as the familiar foot-rule, or it may be as subjective as a personal rating made by one person about another.

III. Origin and Development of Measurement

The origin of measurement goes back farther than historical records. In the most ancient ruins of buildings we find evidence of standards of measurement. Measurement by such standards may not have been measurement in the modern sense, but the ruins show that ancient buildings were erected according to some regular unit. In many cases, if not in all, the ancient units seem to have been in terms of some part of the body. The "foot," it is thought, first appeared in Greece, and the standard was said traditionally to have been derived from the foot of Hercules. Tradition has it that Charlemagne later established the length of his own foot as the standard of measurement for his country.

There is an ancient story of a poor man who, in a thoughtful mood, asked of a wise man, "Why am I poor?" The wise man cut a staff as high as his thigh, made notches in it which were his hand's width apart, gave it to the poor man, and said, "I give you a sceptre of success, a measuring stick; measures rule the world. They go in pairs—the measure of the sandal must match the measure of the foot; so all things are made to measure. With this stick measure what you make; measure well for use. Three loops of cord make it a

balance to weigh what you buy or sell. Set it upright in the sun, and the stick will measure the shadow hours of time—allot then thy tasks. Attune thy life to its circling shadows; when in spring the noon shadows grow long, it is time to plant. Measure your portion and your neighbor's. Make wisely, measure truly, and trade justly, and you will prosper."

The development of measurement in psychology has followed very closely the history of measurement in the other sciences. In most of the sciences we find a stage in which there is much philosophy, but little measurement. We may mention alchemy and its relation to chemistry, or astrology and its relation to astronomy. These two forerunners preceded the development of the two natural sciences and the use of exact measurement. Alchemy and astrology existed at times when very few of the factors involved in the sciences were known. Phenomena were interpreted subjectively, in terms of vague indications. Development of more exact measurement now gives us the precise, almost infallible, predictions of the science of chemistry or of astronomy. When completed, the history of measurement in psychology may read very much like the history of measurement in such sciences as these. A difference at the present time lies in the fact that psychology has not yet progressed so far beyond subjective, introspective types of measurement. In a later chapter, we shall attempt to follow in more detail the history of psychological measurement from the beginning of the development of definite standards.

IV. Value of Measurement

Progress in any line depends on ability to measure exactly. The steam engine could not be created until

8 Value of Measurement in Psychology

man was able to make a piston and cylinder of such exact dimensional relationships that although little steam could escape between them there would remain sufficient room for the piston to move up and down. The automobile had to wait until man could measure to one five-thousandth of an inch. Owing to lack of accurate measuring devices, medicine remained largely speculative almost until modern times; it is only within the last hundred years that medical literature has exhibited anything like scientific accuracy. Exactness of measurement is the best index of the development of a profession.

We may indicate the importance of measurement to progress in psychology by several quotations from men who have contributed much to it:

The history of science is still written chiefly in terms of man's struggle to know and master the world of nature about him. It is a story more dramatic and more significant in the history of the race than the story of sieges and lost battles and diplomatic victories. But significant as the achievement of man may be in the conquest of force and matter, of air and land and water, its telling may, in the years to come, give way before a new epic, the story of man's study and growing knowledge of himself as an individual and in the mass. Of this new story, only a few chapters have as yet been written, in comparison with what may still be before us to know and to write.

For such a study new norms and new techniques must be derived and applied. They are already being set up and, as in all sound scientific progress, they are built upon the disciplines and the knowledge that the world of scholarship has laid at the basis of past progress. And as in the past, all fields of exact knowledge or bold scientific theory were based on measurement, that is, upon mathematics, so in the study of himself and his kind, man seeks the aid of mathematics and

develops bio-chemistry as the handmaid of his newer studies of himself and of all animate things.¹

Science, as it develops, advances through the stages of observation, description, and experimentation to the stages of measurement and calculation. Our object is to consider the methods, and in some small part, the accomplishments of those students of mankind who have studied human individuals for the purpose of basing broad and permanently valuable mathematical generalizations on masses of measurements.²

Psychology had to await the development of the exact and the natural sciences, whose objects are more open to measurement, whose contents are more basic, and whose applications are more useful. And it should be remembered that all the sciences, as we now know them, are comparatively new. The doctrine of the conservation of energy is only about as old as Professor Stumph, the theory of evolution by natural selection about as old as Professor Jastrow. Modern physics and modern genetics are no older than the younger members of this congress. When sciences of earlier origin have made such notable advances during the lifetime of those now living, we may look forward with hopefulness to a corresponding development of psychology within the lifetime of our children.³

Psychology cannot attain the certainty and exactness of the physical sciences, unless it rests on a foundation of experiment and measurement.⁴

There are able psychologists who like to narrate what they think they think, what they feel they feel, what

¹ Harris, J. A., Jackson, C. M., Paterson, D. G., Scammon, R. E.: *The Measurement of Man*, The University of Minnesota Press, Minneapolis, 1930, p. v.

² *Ibid.*, p. 5.

³ Cattell, J. McKeen, *Psychology in America*, the Science Press, New York, 1929, p. 2.

⁴ *Ibid.* (quoting from article on "Mental Tests and Measurements," *Mind*, No. 59, 1890), p. 30.

10 Value of Measurement in Psychology

they imagine they imagine. Those of us who are concerned with quantitative measurements and objective results . . . think . . . that such literary diversions contribute about as much to a science of psychology as similar stories about their rheumatism and other bodily ailments would contribute to a science of pathology.⁵

In psychology, measurement may be considered the master key by which we secure understanding, prediction, and control of behavior. We understand the scholastic failure of the high-school boy by measuring his intelligence; we understand the emotional, high-strung behavior of the hyperthyroid by measuring his metabolic rate; or we understand the automobile accident by measuring the reaction time of the driver. The prediction of behavior through measurement is illustrated by the determination of possible college success through scholastic ability tests, and the determination of future vocational success by vocational aptitude tests. Control of behavior is made easier by measurement, because through it we have a more complete knowledge of those things which we are to control, both in terms of the human factor and in terms of environment which may influence the human factor.

The value of measurement to psychology is indicated also by the part it has played in the development of the various divisions of psychology. Experimental psychology, for example, rests upon the ability to measure psychological factors quantitatively, and it had its origin in the adoption by psychologists of measurements first borrowed from physics. The whole field of psychological testing is in itself the field of measuring human

⁵ *Ibid.* (quoting address of the President of the American Association for the Advancement of Science, Jan. 8, 1926), p. 32.

qualities. Industrial psychology has depended for its success upon the quantitative measurement of vocational aptitudes, and the quantitative study of factors important in the control of employees. Child psychology has depended upon our extension of experimental and quantitative methods to the study of behavior of even the youngest infant. These various spheres will be discussed in more detail as the various types of measurement are taken up in the book. They are mentioned here simply to indicate the value of measurement to the whole of psychological progress. As in other scientific fields, measurements in this one are an index of scientific status, of the accuracy and exactness with which we do things.

V. Problems of Measurement in Psychology

Thorndike once made a statement that whatever exists at all exists in some amount, and whatever exists in amount can be measured. This would lead us to the conclusion that all the realities with which we deal in psychology are at least potentially capable of measurement. Why, then, has measurement in psychology progressed so slowly? The answer to this question is many-sided and has many contributing factors. One contributing element lies in the complexity of the things to be measured; and still another lies in the attitude of people themselves toward allowing the qualities of mankind to be measured. Finally, let us quote one more paragraph from the address of J. McKeen Cattell, a pioneer in the field of psychological measurement:

It may at first sight seem surprising that the sciences, both curious and useful, of matter and energy should have had an earlier origin and a more systematic development than the biological sciences, while psychol-

12 Value of Measurement in Psychology

ogy is only now taking its place among the descriptive sciences and has witnessed but its first beginnings as an applied science. The explanation is partly in the difference in stability and complexity of the objects of the different sciences. Matter is plastic to experiment and measurement; human behavior eludes experimental and quantitative methods. The motions of the solar system since its beginning are less complicated than the play of a child for a day. It is also the case that in the architectonics of science the mathematical and physical sciences are fundamental. Morphology and physiology are based upon physics and chemistry; psychology on all these sciences. The foundations must be laid before we can build the upper stories, in which we may prefer to live and from which there may be a wider outlook.⁶

⁶ Cattell, J. McKeen, *op. cit.*, p. 1.

CHAPTER II

The Instruments of Measurement in Psychology

MEASUREMENT has already been defined as a means of indicating something quantitatively. In order to be able to indicate something quantitatively, we must have some measuring standard or yardstick, as it were, which is generally acceptable. Practically all measuring units in the physical sciences represent standards of length, weight, or whatever else is being measured; these standards are uniformly the same wherever and whenever used. We can readily see, then, that some such units must be established if we are to measure things in the field of psychology. In some instances the psychologist has simply adopted already established units from the physical sciences. He may, for example, be interested in weight measurement as indicative of physical growth in the child. For this he borrows the standard unit of mass or weight, and compares with it the mass or weight of the child which he may have in question. If he is interested in measuring strength, again he borrows a standard from the physical sciences. In measuring mental, emotional, and personality characteristics, the psychologist had to invent units of his own, but in the building up of these new types of units the same principles, in general, hold. The unit still represents a standard with which the thing to

be measured is compared, either directly or indirectly. For many of the psychological qualities, the comparisons must be indirect rather than direct, as we shall see when we discuss the units in more detail.

It seems well, before a discussion of specific psychological tests, to outline the general types of measurement or of units that have been developed. Three main types of measuring devices seem to belong distinctly to the psychologist: (1) psychological tests; (2) rating scales; and (3) questionnaires.

I. Psychological Tests

By the term "test" we shall mean here a means of measuring which involves solving a problem or performing a task with an accomplishment to show for it. The degree or amount of the trait being measured is indicated by the degree of the accomplishment. The accomplishment itself may be graded as to amount accomplished, quality or correctness, or time taken to attain it. In any event, the subject does something and his accomplishment shows the measurement of the trait in question.

Most of these measurements are indirect; that is, they directly measure not the trait or ability, but the results of this trait or ability in action. In other words, in the intelligence test, for instance, intelligence is measured by the psychologist not directly but by the accuracy and amount of performance on something which demands the use of intelligence. We do not know in most instances how to make a direct approach, and even in those in which directness might seem possible it may be impracticable. We may believe, for instance, that intelligence can be directly measured in terms of the number of neurones a person has, but in a living brain this is obviously an impracticable approach to the problem.

Some brief examples of psychological tests may make the meaning of tests somewhat clearer.

Intelligence tests are tasks which necessitate the use of intelligence in their performance. The individual intelligence test involves the performance of varied tasks such, for example, as repeating syllables after an examiner has spoken them, or carrying out commands which are given orally by the examiner, or copying designs from a sheet of paper. An ordinary pencil-and-paper group intelligence test may involve answering questions requiring judgment, working problems in arithmetical reasoning, or answering questions about the meanings of words. Non-verbal intelligence tests for very young or illiterate individuals may involve putting together puzzle blocks into a complete picture, or placing cut-out designs in a form board, or tracing a continuous line through a maze.

A test in *aptitude* may involve speed of tapping to predict possible future speed in typing, or the answering of questions based upon information which must be acquired before a vocational job can be undertaken.

A test of *social intelligence* may involve answering questions about social situations which necessitate the use of social judgment; or answering questions reflecting one's observation of human behavior and human motivations.

A test of *mechanical intelligence* may involve putting together a dozen pieces to make up some mechanical device, or the identification of commonly used mechanical tools.

These will serve as examples of what is meant by a "test" as the term is used in psychological measuring.

Tests used by psychologists may be of two types: *objective* and *subjective*. They are objective when the

performance can be graded, rated, or scored the same by everybody who attempts to grade them. Such tests can be scored by a definite scale which does not necessitate judgment or decision on the part of the person doing the rating at the time the performance is graded. In other words, an objective test is one in which a given performance will produce the same measurement no matter who rates it.

A subjective test, on the other hand, is one in which the judgment or decision of the rater has a great deal to do with the grade or rating which a given performance receives. To illustrate the differences between these, suppose we wish to disclose a person's knowledge of English literature. By an objective method, we might give him a list of 50 definitely true or false statements about books, writers, and movements in the field of English literature, and perhaps in addition, a list of 20 books or poems to be matched with an equal number of authors. Once constructed, such a test could be graded with exactly the same results by everybody who graded it. On the other hand, by a subjective method, we might ask the individual in question to discuss first the importance of Carlyle and then, perhaps, the romantic movement in English literature. It is quite obvious that several people attempting to grade an individual's performance on such a test might disagree in their evaluation of it.

II. Rating Scales

The rating scale is another means of indicating quantitatively the degree to which individuals possess abilities or traits. In this method, however, instead of having the individual perform some task or problem which is to indicate the amount of the trait, the measurement represents the subjective impressions of someone who

judges the amount of the trait or ability possessed by the individual from previous association with him in situations where the trait or ability is supposed to be shown. In other words, ratings represent judgments made about someone. They may be made about oneself or about another, more frequently the latter. The value of ratings depends fundamentally upon the accuracy of the judges who make them, upon the number of judges, and also upon the extent to which the abilities rated lend themselves to observation and evaluation. Several types of rating scales have been developed. These are discussed in Chapter XIX.

III. Questionnaires

A questionnaire represents in general a systematic report of an individual's experiences, attitudes, interests, or beliefs, given by his answers to certain questions. It differs from a psychological test primarily in that it usually asks for a statement about an individual's personal characteristics or his personal beliefs, rather than for the performance of a definite task or problem. It differs from a rating scale in that it usually does not ask for estimates of the degree to which an individual possesses certain traits. It may in some instances, however, overlap the rating scale. Both of these measuring devices may involve the stating of attitudes toward things or of degree of interest in certain things.

Questionnaires have been used as a rule in the study of personality characteristics such as fears, worries, etc., of social attitudes and beliefs, and of vocational interests. Like the rating method they are also used primarily with traits which are not capable of definite measurement by a test method, and again as with the rating method, the value of the questionnaire can be improved by attention

to the definiteness and the objectiveness of the questions. (See Chapters XI and XIX for examples of questionnaires.)

IV. Raw Measures and Relative Measures

Raw measures. Raw measures represent those values which we obtain first and most directly from the unit employed. For example, if we use an intelligence test with 100 questions, the raw measure represents the number answered correctly. Or, in a mechanical aptitude test in which an individual assembles 10 mechanical devices of a dozen pieces each, the raw measure may be the number of the mechanical pieces properly placed. On a graphic rating scale, a raw measure may represent the simple sum of the numerical credits allowed on each part of the rating scale. These raw measures represent, then, the simple added credits which have been assigned to each part of the test or the scale.

If we had only one measuring device, one yardstick for each quality measured, and if each numerical unit were equal to every other numerical unit on the scale, then we should have no need for anything except the raw measures, just as in measuring length we may designate everything in terms of number of feet. In the latter case we are always measuring with the same device, and each foot is equal to every other foot.

Relative measures. In psychological measurement, however, we are not always using the same device. We may use the Army Alpha Test, by which we measure intelligence with a scale extending to 212 points; or the Binet Test, with a scale in terms of months of mental age; or the Thorndike Intelligence Test, with a scale extending over some 500 points. In many of our scales the problem is further complicated by the fact that an

increase of five points on the scale may not mean exactly the same thing at different places on the same scale. Therefore, in order to make our measurements interpretable without a very detailed knowledge of each available measuring device, we resort to the use of *relative measures*. These are derived from the raw measures in such a way as to indicate the ratings of an individual in terms of his relationship to others of a similar group or in relation to others in the group of which he is a part. Most of these relative measures will be discussed in detail in their proper places; they will only be mentioned at this point.

In measurement of the intelligence of children, probably the most commonly used relative measure is the I. Q., or Intelligence Quotient, which translates the raw intelligence measurement into a rating in terms of the child's relationship to other children of his chronological age. In adult measurement probably the most common relative measure used is that of percentile rating, which states the individual's position in the group in which he is measured, assuming that the group is made up of 100 persons. Other means of indicating relative measures are based upon placing the individual in his proper position in relation to a normal distribution curve of the trait being measured, the amount of the trait usually being stated in terms of how far the person diverges from the average or central tendency of the measurement.

V. Reliability and Validity

Every measuring device must possess the characteristics of *reliability* and *validity*. We shall have occasion to refer to these terms many times in the course of our discussions, for the meeting of these standards forms the

“acid test” of our ability to measure human traits and abilities quantitatively.

A test or measuring device possesses reliability if it gives the same result when applied at different times or when applied by different persons. To illustrate with a single physical measurement, a yardstick possesses reliability for measuring length in feet and inches because its use gives the same result for a given distance every time the distance is measured; furthermore, its use gives the same result whether *you* measure the distance or *I* measure it. To apply the principle to a psychological measurement, a test for measuring intelligence possesses reliability if its use gives the same degree of intelligence for a given person upon different applications by different testers. The reliability of a psychological test is ordinarily ascertained by giving the test twice to the same group and comparing the results of the two measurements. To the extent that the results of the two testings are the same the test can be assumed to be reliable. Reliability of psychological tests and measurements depends primarily upon certain characteristics of the measuring instrument, the most important of which are comprehensiveness and representativeness in the content of the test, and objectivity (freedom from dependence upon subjective opinion, judgment, and the like).

A test or measuring device possesses validity when it measures what it purports to measure. Our yardstick is a valid measure of length because it actually measures the physical quality in question. A test designed to measure intelligence is a valid test of intelligence if it actually measures mental ability; it is not a valid test if it measures something else. A test for measuring vocational ability is a valid test of that ability if those rating high on the test actually possess high ability in the voca-

tion and those rating low possess low ability in the vocation. Because of the indirect nature of many psychological tests and measurements, validity has not been an easy standard to attain, and many of the measuring devices studied have fallen short of satisfactory validity. The validity of a psychological test is ordinarily ascertained by comparing the test measurements for a group of persons with some other measurement or indication (criterion) of the trait or ability which the test purports to measure. If the test measurements agree with the criterion, then the test may be assumed to be valid; if they do not, then the test is invalid. For example, the validity of a vocational aptitude test may be ascertained by comparing test results with actual records of vocational success of those who have taken the test.

Correlation. We shall so frequently refer to this term in studies of reliability and validity of psychological tests that it seems wise to define it at this point. *Correlation*, as the term is ordinarily used, refers to a statistical method of determination of the relationship between two things, measurements, or variables. The amount of relationship is indicated by a numerical result, a *coefficient of correlation*, which may have a numerical range from 0 to 1.00. If two things being studied by the correlation method vary in the same direction—that is, if when either is high in amount the other tends to be high—the correlation is positive; if they vary in opposite directions—if when one is high in amount the other tends to be low—the correlation is negative.

The coefficient is *zero* if the two things being studied are not related at all. For example, should we study by the correlation method the relationship between height and intelligence in adults, we should expect the correlation coefficient to be zero, since these two variables are

unrelated and are not found to vary together. The coefficient is 1.00 if the two things being studied vary together in perfect order. For example, if we should compare by correlation scores or ratings on two psychological tests given to a group and should find that the highest person in one was highest in the other also, that second in one was second in the other also, that third in one was third in the other, and so on, the coefficient of correlation would be 1.00. Relationships short of perfect are represented by coefficients of correlation between 0 and 1.00. If the relationship be nearly perfect, the value may be .85 or .90; if only slight, .10 or .20.

Correlation coefficients are a very convenient means of stating many of the relationships we shall have to discuss in considering psychological measurements. They enable us to summarize by one numerical expression what we might otherwise be able to express only by presenting much more detailed data. Test reliability and validity, which we have just discussed, are commonly studied by the correlation method. The relationship between the two applications of a test are "correlated" to indicate reliability. Psychological test scores are "correlated" with the criterion measures to indicate validity. It is difficult to answer the important question that must arise in the reader's mind at this point. What should be the numerical value of coefficients of correlation worked out to indicate reliability and validity? Undoubtedly, they should be as high as possible to attain. Aside from this general statement, we should perhaps simply warn the user of psychological tests against putting too great a dependence on those which show relatively low reliability and validity coefficients.

CHAPTER III

The History of Measurement in Psychology

PRACTICALLY every child in public or private school today has taken a “psychological” test or a mental test—has had some one or more of his mental traits measured. And the adult who has been placed vocationally in the last five years more likely than not has had some test of his mental capacity for performance of his job. This use of measurement seems to have spread in a short time. The term “mental tests” was first used in 1890 by Cattell to designate tasks which he hoped would measure intellectual ability. The first individual intelligence test, that of Binet, appeared in 1905. The first group intelligence tests as we know them today were applied during the World War, less than twenty years ago. And within the twenty years after the first intelligence test was used, psychological testing devices had reached practically the stage of development that we see today.

In any particular development in a broad scientific field we are likely to be doubly impressed—first, we are amazed by the rapidity with which progress takes place once the beginning has been made; and second, we wonder why development did not begin sooner. We are often misled in our judgments of rapidity of development because we cannot see all the development—we do not see the roots which go far back into the history of science. For example, we may see the beginning of intelligence

tests in Binet's test of 1905, yet the roots of development of his test go much farther back. If we wonder why intelligence and other psychological tests did not occur sooner in the history of psychology, let us consider the prerequisites to the development of most applications of science. Except for some few instances of purely accidental discovery, applied fields of a science are developed only when (1) attitudes both of scientists and of the public are conducive to the development, (2) needs for the application exist, and (3) techniques in the science have been refined sufficiently to make the new application possible from a technical standpoint. Intelligence tests had to await thorough acceptance of a belief in individual differences, for without a belief in individual mental differences there was little interest in measurement. Changing attitudes toward the mentally defective and the "peculiar" individual, also, encouraged work on mental tests. The need for these tests had existed for some time in the social problem of dealing with defectives, criminals, and the like. On the eve of the actual development of the first general intelligence test, a need was emphasized to psychologists and educators by difficulties in the guidance of school children who did not progress normally through the grades. Later, probably, the best example of the influence of need is seen in the hastened development of group intelligence tests due to classification problems in the Army during the World War. A survey of the techniques necessary for intelligence test development takes us into the psychological laboratory, to those working on the problems of individual differences, and to the mathematicians from whom psychologists borrowed the many statistical procedures which have been indispensable to the development of testing.

The history of measurement falls into seven periods. These have been designated solely for convenience of discussion, and are not to be thought of as clear-cut in any sense. They merge one with the other, and there is a great amount of overlapping among the types of work represented. They are: (1) Background Period, or Pre-Binet Period; (2) Binet Period—The Individual Intelligence Test; (3) Preliminary Period of Group Mental Tests; (4) World War Period—The Army Tests; (5) Post-War Period—Application of the Army Tests and Development of Similar Tests; (6) Maturing and Stabilizing Period of 1921–1926; (7) Present Period.

If we seem to confine our discussions to intelligence tests, it is because other types of psychological measurement—aptitude tests, personality tests, etc.—have been largely an outgrowth of intelligence tests, taking place in the last two periods.

I. The Background of Intelligence Test Development —The Pre-Binet Period

During this period much was contributed indirectly, if not directly, by experimental psychology. Laboratory activity brought to psychology an objective aspect which the introspectional methods had not produced. Quantitative psychology came into being—measurement became a part of the investigation of psychological entities. At first many of the quantitative aspects were borrowed from physics, so that investigators were in reality as much physicists as psychologists or were individuals who had deserted their original chosen field of physics. Their experiments included such things as physiological studies of the various senses and measurements of thresholds of stimuli for the various senses, sensation levels, speed of

neural impulses, and reaction time. Although today we may not utilize many of the exact measurements with which early experimental psychologists busied themselves, the principles and attitudes established by them have been of immense value in the development of measurements of human traits.

It does not seem out of place to mention a few of the greatest names in this early development of quantitative psychology. The honor of founding quantitative psychology is most frequently given to Gustav Theodor Fechner (1801–1887). His book *Elemente der Psychophysic*, published in 1860, was the starting point for quantitative psychology. In this book Fechner brought together scattered observations from astronomy, physics, and biology, related these to his own elaborate observations in physics, mathematics, and physiology, and placed them all at the service of psychological measurement. Fechner's important contributions to psychology can be named under three headings: a clear expression of Weber's Law (a law regarding the relation between sensation and stimulus); an elaboration of the concept of the threshold as regards stimuli; and the working out of three independent psychophysical methods for measurement of thresholds.

Helmholtz may be mentioned as a second great figure in early quantitative psychology. In many ways he did more to inspire the advance of quantitative investigation than did Fechner. Fechner was much of a philosopher, with not a little of the mystic in his disposition, and his investigations are too often bound up with theories reflecting his mysticism. Helmholtz, on the other hand, was wholeheartedly a scientist and an empiricist. His greatest contributions lie in quantitative investigations in sensory fields of vision and hearing. His investigations

also included measurements of speed of neural impulse and reaction-time experiments.

The last of three great figures we shall mention in this period was Wilhelm Wundt (1832–1920). Wundt seems to have done more than any other scientist in this period to inspire students to work in the field of quantitative psychology. He established the first psychological laboratory at Leipzig in 1878. Here began the extensive sensori-motor measurements which were to be so important in the early attempts to measure intelligence. Wundt's influence is seen in the fact that almost all of the early psychological laboratories in America as well as elsewhere were established by pupils of Wundt, and patterned after his laboratory.

G. Stanley Hall founded in 1883 at the Johns Hopkins University what is usually considered the first American laboratory for psychology. Hall later went to Clark University and was instrumental in the development of experimental psychology there. In 1888 Cattell, a pupil of Wundt, established a psychological laboratory at the University of Pennsylvania; and three years later he left to go to Columbia University, where he established another. E. W. Scripture in 1892 founded a similar laboratory at Yale. This venture, however, did not fare so well as did those of Hall and Cattell, and did not attract so many distinguished students.

During the period covering the fifteen or twenty years before the publication of the first Binet Scale in 1905, there appeared an extensive literature on the study and measurement of individual differences and the use of various sensory and motor tests. These represent the beginning of "mental testing" as such; many of them represent definite attempts at measuring the general intellectual capacity of individuals.

In summarizing the character of this period or group of studies we note the following: First, there was an emphasis on individual differences and their study. The interest in individual differences grew out of work in the psychological laboratory and out of studies in related fields of genetics, anthropology, and education. Second, the attempts to measure individual differences were in terms of elemental qualities or simple sensory or mental processes. For these seemed to the early investigators to be the units of mentality or personality as a whole; they were more objectively and reliably measureable; they seemed to offer a better basis for scientific study; and instruments and materials for measuring many of them were directly to be borrowed from the psychological laboratory. The measurement of complex processes and "general" intelligence did not come until the measurements of the simpler processes had proved inadequate as measures of general ability. Third, a large proportion of the tests used in these early attempts to study mental ability quantitatively were of the sensory or motor type, as tests of sensory acuity, ability to discriminate weights, ability to react quickly to a motor stimulus, tests of strength, etc. These reflected the influence of the early psychological laboratories in which methods of this kind had been worked out. It was expected or hoped that such simple sensori-motor measurements would be indicative of the general ability of the individual. Fourth, this period marked the introduction of statistical procedures for the study of the value of mental measurements.

II. The Binet Period

The Binet Period is that which culminated in the development of the first real intelligence test. Alfred Binet stands as the most important figure in the formulation of

This test, which is today known as the Binet Test or the Binet-Simon Test, after Binet and his coworker. Binet had studied medicine, and through contact with Charcot and others had developed an interest in abnormal psychology. In fact most of his early work dealt with medical or physiological subjects. He seems to have given up abnormal psychology by 1887 and thereafter to have devoted himself to problems more closely allied to educational psychology. He began experimentation in the schools of Paris and its suburbs.

When Binet began this work, unsatisfactory results were being obtained with the "intelligence" or "mental" testing already going on. But failures did not stop the work. Numerous special studies were made of various aspects of the problems which developed. Binet was a most energetic worker in these studies, and there is little doubt that his findings were important in the final evolution of his intelligence scale, which was first published in 1905. In fact, many of the separate tests of his later scales appear in his earlier studies.

A bibliography of Binet's work shows that his early investigations included a study of memory and imagination, with suggested tests for measuring them. In these studies Binet first used the memory for syllables and sentences as a test, a method of measuring later constituting part of the Binet Intelligence Scale. He also reported results in the measuring of suggestibility; investigations of the effect of various physical processes, as eating, on mental work; and study and measurement of attention with a consideration of its relation to intelligence. Probably the two most important writings of Binet previous to the publication of his intelligence scale are an article appearing in 1898 in the *Revue Philosophique* and a volume appearing in 1902 entitled *The*

Experimental Study of Intelligence. The former discusses measurement in individual psychology and sets forth many of Binet's ideas as to how intelligence would have to be measured. Binet recognized that indications of intelligence, to be valuable, must represent measurements, not merely descriptions. He also pointed out in this early article the inadequacy of the simple tests of sensory and motor processes for measuring the sum total of intelligence; he did not at that time suggest the solution of the problem, but he set down some of the main theses on which the work must be developed. He emphasized the importance of standardized conditions where the testing of mental processes is carried out, and he suggested the use of a gradation of tasks as to difficulty in the measurement of mental qualities. Several tests mentioned in this article we find later in the construction of the Intelligence Scale of Binet and Simon. *The Experimental Study of Intelligence* reports an extensive qualitative analysis of the mental abilities of his two daughters, with less extensive observations of responses of a number of other subjects. Many tests later used by Binet appear in this report. From a theoretical standpoint the report is of great importance in its emphasis upon the testing of higher processes—processes stimulated by language and other social stimuli—as against the testing of mere sensory stimuli and the recording of responses of the simpler type, which Binet held inadequate for indicating the individual's intelligence.

In the latter part of 1904 the Minister of Public Instruction appointed a commission, of which Binet was a member, to investigate the problem of backward children in the Paris schools. In the course of the investigation an immediate need for an intelligence scale presented

itself, and it was to meet the emergency that the first intelligence scale was constructed.

By this time Binet had rather clearly conceived the ideas which made his scale so markedly superior to previous attempts to measure intelligence. These ideas may be summarized as follows: (1) Measures of simple sensory and motor processes are not adequate for measuring intelligence; higher, more complex processes must be measured. (2) A wide variety of tasks or tests must be administered to obtain an adequate measurement of intelligence. (3) Conditions under which the tests are given must be standardized and kept constant. (4) The various parts of the test should show gradations in difficulty. (5) Age standards for expressing intelligence are feasible and practical (this idea, however, did not appear in the first scale but was utilized in the revision three years later).

The first of the Binet intelligence scales appeared in 1905. It consisted of 30 tasks of varying degrees of difficulty, arranged in order from easiest to most difficult, the difficulty being determined from trials on normal and subnormal school children. The tasks were all relatively simple and easily administered, and the results relatively easily scored. Examples of tasks 1, 16, and 28 are given below:

1. Visual coördination. Noting the degree of coördination of movement of the head and eyes as a lighted match is passed slowly before the subject's eyes.
16. Giving differences between various pairs of familiar objects recalled in memory: (a) paper and cardboard, (b) a fly and butterfly, and (c) wood and glass.

28. Giving the time that it would be if the large and the small hands of the clock were interchanged at four minutes to three and at twenty minutes after six. A much more difficult test is given those who succeed in the inversion; namely, to explain the impossibility of the precise transposition indicated.

The subsequent revisions of the scale differ chiefly in that the number of tests was increased and the tests were arranged into groups according to age, so that intelligence could be expressed as a Mental Age dependent on the group of tests reached in the scale (reaching the tests of the 8-year level, for example, signified a Mental Age of 8.) A more detailed discussion of the final Binet scale appears in Chapter IV.

III. Preliminary Period of Group Mental Tests

After the reliability and validity of the Binet method of intelligence testing were established, the need for some group method of testing was soon felt. If a group method could be developed it would conserve time in testing and greatly extend the usefulness of testing. The need for a sorting device which would quickly classify according to ability a million or more men, recruited during the World War, greatly hastened the final evolution of a practical group test of intelligence.

Prior to the war sporadic instances of group testing occurred. Many of these were purely adaptations to group administration of some of the tests of the Binet series or some of its modifications. Reuel H. Sylvester describes an early instance of such testing:

The writer recalls a situation in 1916, in which the urge toward group testing was felt most keenly. This was before the hastened development of group methods under pressure of necessity for marshalling

the Nation's man-power in the organization of our citizenry into armies for service in the World War. The Medical School of the State University of Iowa had been invited to make a comprehensive study of the pupils in the public schools of Wapello, Iowa. Each pupil was given medical examinations by specialists in a dozen fields of medicine, dentistry, and school hygiene. Experts on heredity and environmental factors studied the family and community aspects of each child's life. The writer was assigned the task of measuring all children as to intelligence and of diagnosing the mental deficiencies of those presenting problems. At that time our only available measuring devices were the Stanford-Binet Scales. To apply any of these to some 500 children was a task entirely beyond our resources of personnel and time.

We met the situation by adapting several of the Yerkes Point scale tests to group application. After these had been given and scored, clinical psychologists took the children one at a time, cleared up doubtful responses that had been written in the group testing, and applied the remaining tests of the Point Scale. Thus fairly accurate scores were secured.

Results from the Wapello survey were so satisfactory that the method of group testing was applied in several other school systems in Iowa. At Council Bluffs especially valuable results were obtained. There an adaptation of the Stanford-Binet Scale was used. Superintendent Theodore Saam followed up and refined the method and used it extensively later. These early efforts are reported as representative of such attempts at group mental testing of that early period. With the exception of parts of Otis' tests no actual products of that Period remain in use today, but those beginnings revealed the possibilities of the method and prepared the way for the miraculously rapid formulation of the Army Tests in 1917.¹

¹ Sylvester, Reuel H., "Group Mental Tests" in *Clinical Psychology*, edited by Robert A. Brotemarkle, University of Pennsylvania Press, Philadelphia, 1931.

During this period Whipple published an important book on tests, many of which were suitable for group testing.² A paragraph from his introduction emphasizes the growing interest in tests during this period:

One need not be a close observer to perceive how markedly the interest in mental tests has developed during the past few years. Not very long ago attention to tests was largely restricted to a few laboratory psychologists; now tests have become objects of attention for many workers whose primary interest is in education, social service, medicine, industrial management and many other fields in which applied psychology promises valuable returns.

IV. The World War Period—The Army Tests

During this period psychological-test development and application received the greatest impetus that they have received in their whole history. During the brief period of about eighteen months a psychological examining division was established in the Army, the many forms of psychological tests developed for Army use were constructed, trials and standardization of the test material were made on limited groups, the final tests were administered to almost two million men, and group testing of mental abilities was demonstrated to be as feasible as and in many ways more practical than the individual testing methods which had originated with the work of Binet.

The Army testing work had its origin in a meeting of a group of experimental psychologists at Harvard University. Upon the declaration of war by the United States on April 6, a session was arranged for discussion of the relations of psychology to national defense and for con-

² Whipple, G. M., *Manual of Mental and Physical Tests. Simpler Processes*, second ed., Warwick and York, Inc., Baltimore, 1914.

sideration of ways that psychology could be applied in the solution of military problems. National scope was given to the psychologists' interest when the American Psychological Association joined the more local group.

Efforts of the psychologists who were working on the problem crystallized toward a plan for applying "psychological" or mental tests to the recruits in the Army as an aid in classifying them for training or occupational duties, and in arriving at estimates of their probable value to the service. Within two weeks the president of the American Psychological Association, then Dr. Robert M. Yerkes, presented the plans of the Association for psychological service to the National Research Council of the National Academy of Sciences. The favorable reception of the plan by the National Research Council gave it more official sanction and undoubtedly did much to insure its final adoption by the War Department. A tentative plan for the psychological examination of recruits was submitted to the Surgeon General of the Army on May 1. It met with official favor, but some weeks passed before official machinery was put into action. During the following three months a committee headed by Dr. Yerkes developed a great deal of the test material later used, and carried out unofficial trials of their material in certain Army and Navy stations. Early in August Dr. Yerkes was appointed to the Army with a rank of major to organize and direct psychological examinations for the Medical Department of the Army. The work was now officially begun and, after sufficient trial, the psychological tests were extended to include all Army recruits.

With this chronological review of the Army work let us examine briefly the task which faced those engaged in it. Their general problem was that of preparing an ade-

quate test for measuring the mental level of large groups of men at the same time. They could not utilize the Binet methods, since these necessitated individual examination of each recruit—a method obviously too time-consuming for use in testing upwards of a million and a half men. In addition to being applicable to large numbers of men in a short time, the test set up had to meet certain other requirements—it must not depend upon specific school information, since many of the examinees had had little formal schooling; it should require a minimum of writing by the subject; it should be capable of measuring over a wide range of abilities, so that all men could be measured, from the mentally defective to those of highest ability; it should be easily and objectively scored and rated; and a number of equivalent forms of the test should be constructed to prevent coaching.

The preliminary experimentation and trials of material led to the development of two new types of tests—one, the Army Alpha Test, for persons who could read and write the English language; and another, the Army Beta Test, for persons who could not read or write. A form of each of these is reproduced at the end of this chapter.

In addition to the two new group tests developed, the psychological division made considerable use of individual tests where doubtful results were obtained by group methods. For the individual tests the Binet Test or a modification of it was used with English-speaking or literate individuals, and a scale of manual performance tests³ for illiterates or foreigners.

Intelligence ratings in the Army were usually translated into letter grades, *A* signifying high scores on the test, *B* fairly high, *C*+ somewhat above average, *C* average, *C*—

³ See Chapter IV for an outline of this test.

below average, *D* low, and *D—* and *E* very low. Table I shows the Army test score equivalents for these letter ratings and the percentage of white soldiers receiving each letter grade.

TABLE I
ARMY TEST RATINGS⁴

Letter Grade	<i>Per Cent of White Soldiers</i>				
	<i>Range of Scores Corresponding to Letter Ratings</i>				
	<i>Receiving Grade</i>	<i>Alpha Test</i>	<i>Beta Test</i>	<i>Performance Test</i>	<i>Binet Mental Age</i>
A	4	135-212	100-118	260-311	18.0-19.5
B	8	105-134	90-99	240-259	16.5-17.9
C+	15	75-104	80-89	215-239	15.0-16.4
C	25	45-74	65-79	190-214	13.0-14.9
C—	24	25-44	45-64	150-189	11.0-12.9
D	17	15-24	20-44	90-149	9.5-10.9
D— & E	7	0-14	0-19	0-89	0.0- 9.4

As an immediate result of the Army tests, about 8,000 men were recommended for discharge from the Army for mental deficiency, about 10,000 were assigned to labor battalions or other services requiring only low-grade intelligence, and about 10,000 were recommended for special battalions for special training or for further observation because of their low ability. For higher grade groups much use was made of the test results in assignments to training groups, in vocational placements, and in the development of officer material.

The more far-reaching results of Army testing reach into all the subsequent developments of group testing methods for measuring human traits and abilities, examples of which are to be considered throughout this text.

⁴ National Academy of Sciences, *Memoirs*, Vol. XV, p. 195.

V. The Post-War Period—Application of the Army Tests and Development of Similar Tests

This period, which covered roughly the two years following the World War, was characterized by the spasmodic application of the Army tests to every conceivable kind of group. As is likely to be true where there is over-enthusiasm for a new device, much chaos marked these early trials. Often too much confidence was placed in quickly carried out testing programs; many persons attributed values to the tests that they were never designed to have; and too often the findings of scientific investigators were misinterpreted and given fantastic meanings by lay interpreters and popular publicity. But many of these early studies following the development of the Army tests led to the great expansion of psychological testing which has taken place in the last fifteen to twenty years.

In the development of new tests, the Army testing work bore fruit first in the construction of group tests for use in schools and colleges. One of the earliest of these was a group intelligence test published by Otis. Otis had, in fact, been working on testing material before the war and had contributed many of his ideas and questions to the Army tests. His first academic scale was published in 1918 shortly before the close of the war.⁵ Some forms of the Otis Intelligence Tests are still widely used in measuring intelligence in high school. Another group test for measuring intelligence which appeared very early was a scale by S. L. and L. W. Pressey. Their test was designed for high schools, and though it is not widely in use today, it was used in many of the early school surveys.

⁵ See Chapter V for a discussion of one of Otis's tests.

In 1919 appeared the Haggerty Delta 1 and Delta 2 Tests. The former, largely non-linguistic and similar to the Army Beta, was designed for children of the primary grades; the latter, more like the Army Alpha, was for children of upper elementary grades. Both scales were worked out in connection with mental surveys which the author conducted in Virginia schools.

A year later the National Intelligence Tests appeared. These tests were the outcome of an extensive study of test material by a committee consisting of Haggerty, Terman, Thorndike, Whipple, and Yerkes, the study having been financed by a grant from the General Education Board. Most of the men who worked out these tests had been in the psychological service during the war, and their recent experience was brought to bear in the development of these group tests for academic use. Two scales, *A* and *B*, with two equivalent forms each, were constructed, and by giving the tests to large numbers of children, norms for each age and grade were established. The tests are suitable for the elementary grades. They have been used probably more widely than any other single test for the elementary school.

These examples represent important developments during the period immediately following the Army testing. There were many other studies not mentioned in this brief sketch, and work on some of the important tests appearing later, as the Thorndike Intelligence Examination for College Students, was begun during this period.

VI. The Maturing and Stabilizing Period of 1921-1926

Intelligence testing now advanced to practically its present stage and settled down to its real effectiveness. The period is marked primarily by the following features: (1) New tests in the field of group mental testing con-

tinued to be developed. Particularly was there an extension of the group testing methods to the construction of new tests at the upper and lower ends of the testing scale. College tests and kindergarten or preschool tests, slower to appear than the general adult test and the intermediate school test, began to appear. (2) Much fundamental work was done on the validation of the new group tests of mental ability. These studies firmly established the group method of measuring ability, confirming for the newer tests in the schools what had been found true in the Army testing. (3) Those measuring instruments which stood the tests of validity and reliability were standardized so that ratings on them could be properly interpreted in terms of relative abilities. (4) Many studies of the uses and applications of intelligence tests were made in the schools and, to a more limited extent, in other organizations interested in measuring human traits. Such studies established the use of mental tests on a scientific basis and pointed the way to abandonment of some of the earlier sweeping claims made for mental tests. (5) The methods and techniques of the newly developed intelligence tests began to be applied in developing other measuring instruments. The "short-answer" or "psychological" type of question was early adapted to measuring achievement, marking the beginning of the objective "short-answer" type of classroom examination in the various school subjects. The methods of the general mental test were also early adapted to the working out of more specialized tests, such as the special aptitude tests for vocations and occupations, or the special ability test exemplified by the tests for measuring mechanical aptitude. Finally, there was a growing interest in the measurement of personality and character traits.

VII. The Present Period

The present period of psychological testing, at least for intelligence testing, is one of matured application in which group mental testing has recently changed but little. In academic uses, mental tests have become rather firmly established, their true values are generally understood, and their limitations have been fairly well determined. In other fields, outside the schools, mental testing is not yet on quite so firm a footing; but the last ten years have witnessed a marked increase in the utilization of the methods of psychological testing in such fields as those of employment and personnel work. Newest present-day developments in psychological measurement are those in personality and character testing; in fact, this chapter in the history of our subject is yet far from completed.


We shall not attempt here to sketch in any detail these last periods of the history of psychological testing. The rest of this book will tell their story.

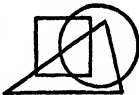
ARMY ALPHA TEST ⁶

TEST 1

1. ○ ○ ○ ○ ○

2. ① ② ③ ④ ⑤ ⑥ ⑦ ⑧ ⑨

3. 

4. 

5. ○ ○ ○ Yes No

6. ○ ○ ○ ○ ○

7. **ABCDEFGHIJKLMN O P**

8. ○ ○ ○ *MILITARY GUN CAMP*

9. **34-79-56-87-68-25-82-47-27-31-64-93-71-41-52-99**

10.

--	--	--	--	--

11.

7F	△ 4	○ 3	△ 5A	○ 8	□ 2	△ 6	○ 9B	□ 3
----	-----	-----	------	-----	-----	-----	------	-----

12. 1 2 3 4 5 6 7 8 9

Score.

1	
2	
3	
4	
5	
6	
7	
8	
T	

⁶ Reproduced by permission of the National Academy of Sciences.

TEST 2

Get the answers to these examples as quickly as you can.
Use the side of this page to figure on if you need to.

- | | | | | |
|---------|---|---|---|---------------|
| SAMPLES | { | 1 | How many are 5 men and 10 men? | Answer (15) |
| | | 2 | If you walk 4 miles an hour for 8 hours, how far do you walk? | Answer (12) |
-
- 1 How many are 20 boats and 9 boats? Answer ()
 - 2 If you save \$4 a month for 9 months, how much will you save? Answer ()
 - 3 If 64 men are divided into squads of 8, how many squads will there be? Answer ()
 - 4 Mike had 11 cigars. He bought 3 more and then smoked 8. How many cigars did he have left? Answer ()
 - 5 A company advanced 6 miles and retreated 2 miles. How far was it then from its first position? Answer ()
 - 6 How many hours will it take a truck to go 48 miles at the rate of 3 miles an hour? Answer ()
 - 7 How many cigars can you buy for \$1.09 at the rate of 2 for 5 cents? Answer ()
 - 8 A regiment marched 40 miles in five days. The first day they marched 9 miles, the second day 6 miles, the third 10 miles, the fourth 7 miles. How many miles did they march the last day? Answer ()
 - 9 If you buy 2 packages of tobacco at 7 cents each and a pipe for 75 cents, how much change should you get from a two-dollar bill? Answer ()
 - 10 If it takes 5 men 4 days to dig a 20' foot drain, how many men are needed to dig it in half a day? Answer ()
 - 11 A dealer bought some mules for \$1,200. He sold them for \$1,500, making \$50 on each mule. How many mules were there? Answer ()
 - 12 A rectangular bin holds 500 cubic feet of hme. If the bin is 10 feet long and 6 feet deep, how wide is it? Answer ()
 - 13 A recruit spent one-eighth of his spare change for post cards and twice as much for a box of letter paper, and then had \$2.00 left. How much money did he have at first? Answer ()
 - 14 If $5\frac{1}{2}$ tons of bark cost \$35, what will $3\frac{1}{2}$ tons cost? Answer ()
 - 15 A ship has provisions to last her crew of 400 men 6 months. How long would it last 1,600 men? Answer ()
 - 16 If an aeroplane goes 300 yards in 10 seconds, how many feet does it go in a fifth of a second? Answer ()
 - 17 A U-boat goes 6 miles an hour under water and 20 miles an hour on the surface. How long will it take to cross a 100-mile channel, if it has to go three fifths of the way under water? Answer ()
 - 18 If 214 squads of men are to dig 5,992 yards of trench, how many yards must be dug by each squad? Answer ()
 - 19 A certain division contains 6,000 artillery, 18,000 infantry, and 1,000 cavalry. If each branch is expanded proportionately until there are in all 24,200 men, how many will be added to the artillery? Answer ()
 - 20 A commission house which had already supplied 1,897 barrels of apples to a cantonment delivered the remainder of its stock to 38 mess halls. Of this remainder each mess hall received 45 barrels. What was the total number of barrels supplied? Answer ()

TEST 3

This is a test of common sense. Below are sixteen questions. Three answers are given to each question. You are to look at the answers carefully, then make a cross in the square before the best answer to each question, as in the sample:

SAMPLE { Why do we use stoves? Because
☐ they look well
☒ they keep us warm
☐ they are black

Here the second answer is the best one and is marked with a cross. Begin with No. 1 and keep on until time is called.

- | | |
|--|--|
| <p>1 Cotton fibre is much used for making cloth because
 <input type="checkbox"/> it grows all over the South
 <input type="checkbox"/> it can be spun and woven
 <input type="checkbox"/> it is a vegetable product</p> <p>2 Thermometers are useful, because
 <input type="checkbox"/> they regulate the temperature
 <input type="checkbox"/> they tell us how warm it is
 <input type="checkbox"/> they contain mercury</p> <p>3 Why are doctors useful? Because they
 <input type="checkbox"/> understand human nature
 <input type="checkbox"/> always have pleasant dispositions
 <input type="checkbox"/> know more about diseases than others</p> <p>4 Why ought a grocer to own an automobile? Because
 <input type="checkbox"/> it is useful in his business
 <input type="checkbox"/> it uses rubber tires
 <input type="checkbox"/> it saves railroad fare</p> <p>5 A machine gun is more deadly than a rifle, because it
 <input type="checkbox"/> was invented more recently
 <input type="checkbox"/> fires more rapidly
 <input type="checkbox"/> can be used with less training</p> <p>6 Why is the telephone more useful than the telegraph? Because
 <input type="checkbox"/> it gets a quicker answer
 <input type="checkbox"/> it uses more miles of wire
 <input type="checkbox"/> it is a more recent invention</p> <p>7 Why is wool better than cotton for making sweaters? Because
 <input type="checkbox"/> wool is cheaper
 <input type="checkbox"/> it is warmer
 <input type="checkbox"/> it wears longer</p> <p>8 Why is New York larger than Boston? Because
 <input type="checkbox"/> it has more railroads
 <input type="checkbox"/> it has more millionaires
 <input type="checkbox"/> it is better located</p> | <p>9 Every soldier should be inoculated against typhoid fever, because
 <input type="checkbox"/> many men have typhoid fever
 <input type="checkbox"/> the doctors insist on it
 <input type="checkbox"/> it prevents epidemics</p> <p>10 Theatres are useful institutions because
 <input type="checkbox"/> they employ actors
 <input type="checkbox"/> they afford a method of relaxation
 <input type="checkbox"/> they give the rich a chance to spend their money</p> <p>11 A train is harder to stop than an automobile because
 <input type="checkbox"/> it is longer
 <input type="checkbox"/> it is heavier
 <input type="checkbox"/> the brakes are not so good</p> <p>12 Why is winter colder than summer? Because
 <input type="checkbox"/> the sun shines obliquely upon us in winter
 <input type="checkbox"/> January is a cold month
 <input type="checkbox"/> there is much snow in winter</p> <p>13 Many schools are closed in summer, so that
 <input type="checkbox"/> the teachers may have a vacation
 <input type="checkbox"/> the children shall not be indoors in hot weather
 <input type="checkbox"/> the schoolhouses may be repaired</p> <p>14 If a drunken man is quarrelsome and insists on fighting you, it is usually better to
 <input type="checkbox"/> knock him down
 <input type="checkbox"/> call the police
 <input type="checkbox"/> leave him alone</p> <p>15 Why are electrical engineers highly paid? Because
 <input type="checkbox"/> their ability is much in demand
 <input type="checkbox"/> they have a college education
 <input type="checkbox"/> they work long hours</p> <p>16 Aeroplanes failed for many years because
 <input type="checkbox"/> they were too heavy
 <input type="checkbox"/> the materials cost too much
 <input type="checkbox"/> the motor was not perfected</p> |
|--|--|

Go to No. 9 above

TEST 4

If the two words of a pair mean the same or nearly the same, draw line under same. If they mean the opposite or nearly the opposite, draw line under opposite. If you cannot be sure, guess. The two samples already marked as they should be.

SAMPLES	{ good—bad	<u>same—opposite</u>
	{ little—small	<u>same—opposite</u>
1	high—low	same—opposite
2	slow—fast	same—opposite
3	large—great	same—opposite
4	danger—safety	same—opposite
5	genuine—real	same—opposite
6	choose—select	same—opposite
7	fault—virtue	same—opposite
8	similar—different	same—opposite
9	jealousy—envy	same—opposite
10	sacred—profane	same—opposite
11	conquer—subdue	same—opposite
12	vanity—conceit	same—opposite
13	allure—attract	same—opposite
14	waste—conserve	same—opposite
15	deride—ridicule	same—opposite
16	censure—praise	same—opposite
17	illustrious—exalted	same—opposite
18	agitate—excite	same—opposite
19	haggard—gaunt	same—opposite
20	con—pro	same—opposite
21	eminent—distinguished	same—opposite
22	conspicuous—prominent	same—opposite
23	depressed—elated	same—opposite
24	orifice—aperture	same—opposite
25	erudite—scholarly	same—opposite
26	recline—stand	same—opposite
27	degenerate—deteriorate	same—opposite
28	martial—civil	same—opposite
29	nonchalance—anxiety	same—opposite
30	torpor—stupor	same—opposite
31	comprehensive—restricted	same—opposite
32	latent—hidden	same—opposite
33	node—knot	same—opposite
34	celestial—terrestrial	same—opposite
35	carnivorous—herbivorous	same—opposite
36	urbanity—civility	same—opposite
37	proclivity—inclination	same—opposite
38	putrid—fetid	same—opposite
39	impecunious—opulent	same—opposite
40	choleric—phlegmatic	same—opposite

TEST 5

The words A EATS COW GRASS in that order are mixed up and don't make a sentence, but they would make a sentence if put in the right order: A COW EATS GRASS, and this statement is true.

Again, the words HORSES FEATHERS HAVE ALL would make a sentence if put in the order ALL HORSES HAVE FEATHERS, but this statement is false.

Below are twenty-four mixed-up sentences. Some of them are true and some are false. When I say "go," take these sentences one at a time. Think what each would say if the words were straightened out, but don't write them yourself. Then, if what it would say is true, draw a line under the word "true"; if what it would say is false, draw a line under the word "false." If you can not be sure, guess. The two samples are already marked as they should be. Begin with No. 1 and work right down the page until time is called.

SAMPLES { a eats cow grass..... true. false
 { horses feathers have alltrue..false

- | | | |
|--|---------------|----|
| iron heavy is | ..true..false | |
| chairs sit are to on | ..true..false | |
| Alaska in cotton grows.. . . . | ..true..false | |
| happy is man sick always a... .. | ..true..false | |
| wood eat and good to are coal.. . . . | ..true..false | |
| Germany of Wilson king is England and .. . | ..true..false | |
| day it snow does every not | ..true..false | |
| war in are useful aeroplanes..... | ..true..false | |
| sounds people some loud annoy. | ..true..false | |
| 10 thunders rains when it always it ... | ..true..false | 10 |
| 11 food is tobacco as valuable a not ... | ..true..false | 11 |
| 12 trees roses sea and in grow the .. | ..true..false | 12 |
| 13 pole north equator mile one from is the the .. | ..true..false | 13 |
| 14 a battle in racket very tennis useful is .. | ..true..false | 14 |
| 15 made cloth wool cotton and is from.. . | ..true..false | 15 |
| 16 seldom forever good lasts luck .. | ..true..false | 16 |
| 17 a ocean cross minutes few can boat the in a .. | ..true..false | 17 |
| 18 seldom birds' diamonds nests are in found .. | ..true..false | 18 |
| 19 love we wrong those us always who... .. | ..true..false | 19 |
| 20 to aid deep great snow a military manoeuvres is.. | ..true..false | 20 |
| 21 never man the show the deeds.. . . . | ..true..false | 21 |
| 22 always is not a a stenographer bookkeeper.. . | ..true..false | 22 |
| 23 never who heedless those stumble are..... | ..true..false | 23 |
| 24 people enemies arrogant many make..... | ..true..false | 24 |

TEST 6

2 4 6 8 10 12 14 . 16...
 9 8 7 6 5 4 3 . 2..
 2 2 3 3 4 4 5 . 5 ...
 1 7 2 7 3 7 . 4 .. 7..

Look at each row of numbers below, and on the two dotted lines write the two numbers that should come next

2	3	4	5	6	7
10	15	20	25	30	35
8	7	6	5	4	3
6	9	12	15	18	21
5	9	13	17	21	25
8	1	6	1	4	1
25	25	21	21	17	17
1	2	4	8	16	32
4	5	8	9	12	13
8	8	6	6	4	4
19	16	14	11	9	6
3	4	6	9	13	18
12	14	13	15	14	16
29	28	26	23	19	14
18	14	17	13	16	12
16	8	4	2	1	$\frac{1}{2}$
15	16	14	17	13	18
1	4	9	16	25	36
21	13	16	15	12	10
3	6	8	16	18	36

TEST 7

SAMPLES	sky—blue. grass—table <u>green</u> warm big
	fish—swims. man—paper time <u>walks</u> girl
	day—night. white—red <u>black</u> clear pure

In each of the lines below, the first two words are related to each other in some way. What you are to do in each line is to see what the relation is between the first two words, and underline the word in heavy type that is related in the same way to the third word. Begin with No. 1 and mark as many sets as you can before time is called

1	finger—hand . toe—box foot doll coat	1
2	sit—chair . sleep—book tree bad see	2
3	skirts—girl trousers—boy hat vest coat	3
4	December—Christmas November—month Thanksgiving December early	4
5	above—top: below—above bottom sea hang	5
6	spoon—soup . fork—knife plate cup meat	6
7	bird—song man—speech woman boy work	7
8	corn—horse bread—daily flour man butter	8
9	sweet—sugar sour—sweet bread man vinegar	9
10	devil—bad. angel—Gabriel good face heaven	10
11	Edison—phonograph . Columbus—America Washington Spain Ohio	11
12	cannon—rifle . big—bullet gun army little	12
13	engineer—engine. driver—harness horse passenger man	13
14	wolf—sheep cat—fur kitten dog mouse	14
15	officer—private. command—army general obey regiment	15
16	hunter—gun fisherman—fish net bold wet	16
17	cold—heat ice—steam cream frost refrigerator	17
18	uncle—nephew. aunt—brother sister niece cousin	18
19	framework—house skeleton—bones skull grace body	19
20	breeze—cyclone . shower—bath cloudburst winter spring	20
21	pitcher—milk vase—flowers pitcher table pottery	21
22	blonde—brunette . light—house electricity dark girl	22
23	abundant—cheap . scarce—costly plentiful common gold	23
24	polite—impolite : pleasant—agreeable disagreeable man face	24
25	mayor—city . general—private navy army soldier	25
26	succeed—fail. praise—lose friend God blame	26
27	people—house bees—thrive sting hive thick	27
28	peace—happiness . war—grief fight battle Europe	28
29	a—b . c—e b d letter	29
30	darkness—stiffness light—moonlight sound sun window	30
31	complex—simple . hard—brittle money easy work	31
32	music—noise harmonious—hear accord violin discordant	32
33	truth—gentleman . lie—rascal live give falsehood	33
34	blow—anger : caress—woman kiss child love	34
35	square—cube. circle—line round square sphere	35
36	mountain—valley . genius—idiot write think brain	36
37	clock—time . thermometer—cold weather temperature mercury	37
38	fear—anticipation regret—vain memory express resist	38
39	hope—cheer despair—grave repair death depression	39
40	dismal—dark . cheerful—laugh bright house gloomy	40

TEST 8

Notice the sample sentence—

People hear with the eyes ears nose mouth

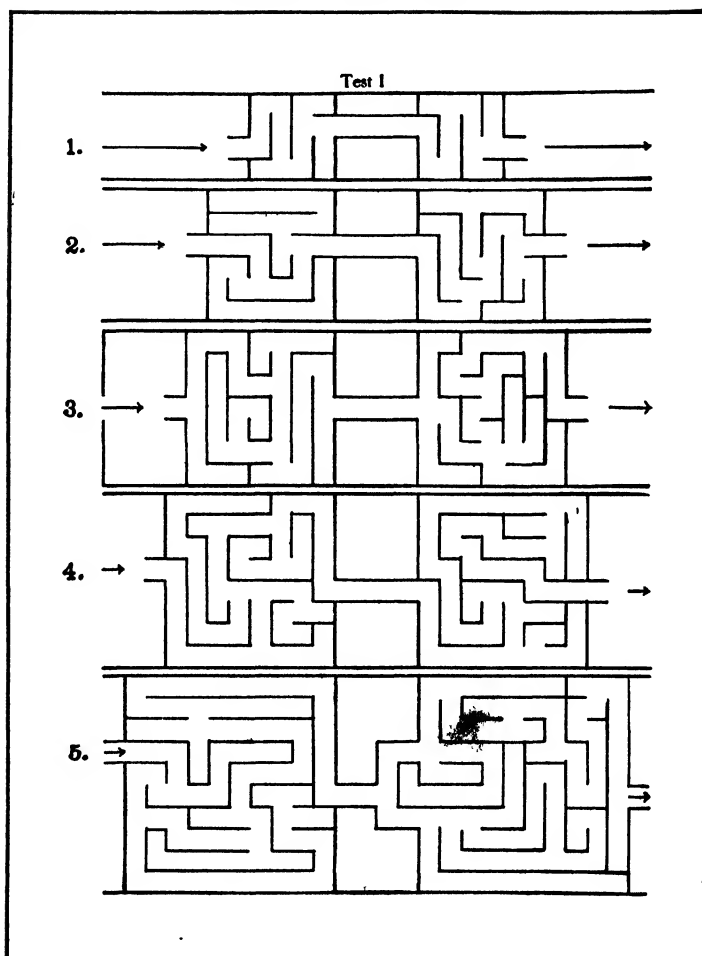
The correct word is ears, because it makes the truest sentence

In each of the sentences below, you have four choices for the last word. Only one of them is correct. In each sentence draw a line under the one of these four words which makes the truest sentence. If you can not be sure, guess. The two samples are already marked as they should be

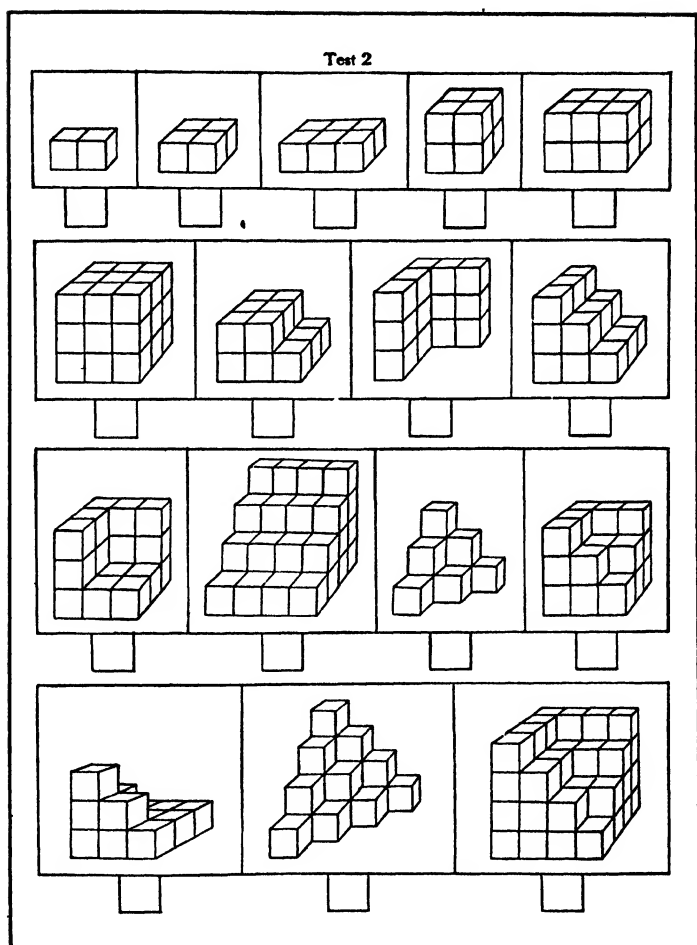
SAMPLES { People hear with the eyes ears nose mouth
{ France is in Europe Asia Africa Australia

- 1 The pitcher has an important place in tennis football baseball handball 1
- 2 Cribbage is played with rackets mallets dice cards 2
- 3 The Holstein is a kind of cow horse sheep goat 3
- 4 The most prominent industry of Chicago is packing brewing automobiles flour 4
- 5 The topaz is usually red yellow blue green 5
- 6 The Plymouth Rock is a kind of horse cattle granite fowl 6
- 7 Irving Cobb is famous as a baseball player actor writer artist 7
- 8 Clothing is made by Smith & Wesson Kuppenheimer B. T. Babbitt Swift & Co 8
- 9 Carrie Chapman Catt is known as a singer writer nurse suffragist 9
- 10 "The flavor lasts" is an "ad" for chewing gum drink health food fruit 10
- 11 Timothy is a kind of corn rye wheat hay 11
- 12 Kale is a fish lizard vegetable snake 12
- 13 The U. S. Naval Academy is at West Point Annapolis New Haven Ithaca 13
- 14 Rio Janeiro is a city of Spain Argentina Portugal Brazil 14
- 15 Emeralds are obtained from elephants mines oysters reefs 15
- 16 John Sargent is famous as a sculptor author painter poet 16
- 17 The iguana is a reptile bird fish insect 17
- 18 The clavicle is in the shoulder head abdomen neck 18
- 19 Karo is a patent medicine disinfectant tooth paste food product 19
- 20 Eucalyptus is a machine tree drink fabric 20
- 21 The carbine is a kind of pistol cannon musket sword 21
- 22 The multigraph is a kind of typewriter pencil copying-machine phonograph 22
- 23 Magenta is a fabric drink food color 23
- 24 The piccolo is used in music stenography book-binding lithography 24
- 25 Cambric is a dance fabric food color 25
- 26 The author of "Treasure Island" is Poe Stevenson Kipling Hawthorne 26
- 27 Blackstone is most famous in law literature science religion 27
- 28 The spark plug belongs in the crank case manifold carburetor cylinder 28
- 29 The Bartlett is a kind of fruit fish fowl cattle 29
- 30 Kelvin was most famous in politics war science literature 30
- 31 Little Nell appears in Vanity Fair Romola The Old Curiosity Shop Henry IV 31
- 32 The number of a Papuan's legs is two four six eight 32
- 33 Arson is a term used in medicine law theology pedagogy 33
- 34 The allo is used in fishing farming hunting athletics 34
- 35 A puck is used in tennis football hockey golf 35
- 36 Dewey defeated the Spanish fleet in Newport News Boston Harbor China Sea Manila Bay 36
- 37 The volt is used in measuring electricity wind power rainfall water power 37
- 38 The Packard car is made in Detroit Buffalo Toledo Flint 38
- 39 The Cooper Hewitt lamp used the vapor of gasoline mercury tungsten alcohol 39
- 40 A regular five-sided figure is scalene rhomboid equilateral elliptical 40

ARMY BETA TEST ⁷



⁷ Reproduced by permission of the National Academy of Sciences.



Test 3

1.

X	X	X	X	X	X	X					
---	---	---	---	---	---	---	--	--	--	--	--

2.

X		X		X		X		X					
---	--	---	--	---	--	---	--	---	--	--	--	--	--

3.

X	O	X	O	X	O	X	O	X	O				
---	---	---	---	---	---	---	---	---	---	--	--	--	--

4.

X	X		X	X		X	X		X	X			
---	---	--	---	---	--	---	---	--	---	---	--	--	--

5.

X	O		X	O		X	O		X	O			
---	---	--	---	---	--	---	---	--	---	---	--	--	--

6.

X	X	O	X	X	O	X	X	O	X	X	O		
---	---	---	---	---	---	---	---	---	---	---	---	--	--

7.

O	O	X	X	O	O	X	X	O	O	X	X		
---	---	---	---	---	---	---	---	---	---	---	---	--	--

8.

X	X	O	O	O	X	X	O	O	O	X	X	O	O				
---	---	---	---	---	---	---	---	---	---	---	---	---	---	--	--	--	--

9.

X	O	X		X	O	X		X	O	X		X	O	X			
---	---	---	--	---	---	---	--	---	---	---	--	---	---	---	--	--	--

10.

X	X	O	X	O	X	X	O	X	O	X	X	O	X	O			
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	--	--	--

11.

X	O	X	X	O	X	X	X	O	X	O	X	X	X	O	X	X	X	O							
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	--	--	--	--	--	--	--

12.

X	X	X	X	O	O	X	X	O	X	X	X	X	O	O	O	X	X	O								
---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	--	--	--	--	--	--	--	--

Test 4

5
U

6
o

7
^

8
X

9
=

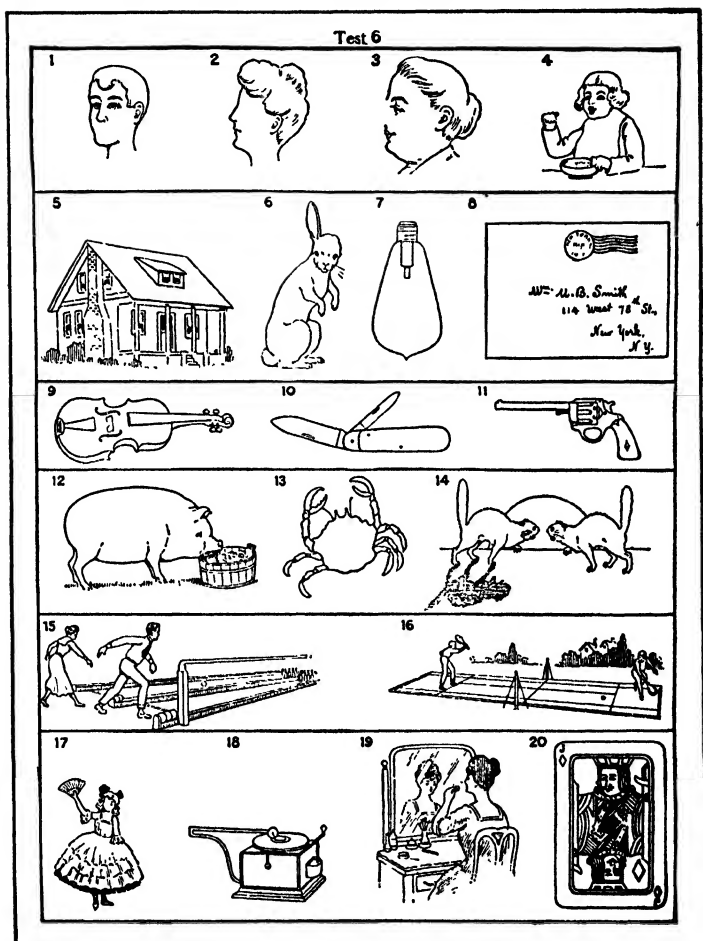
1.

3.

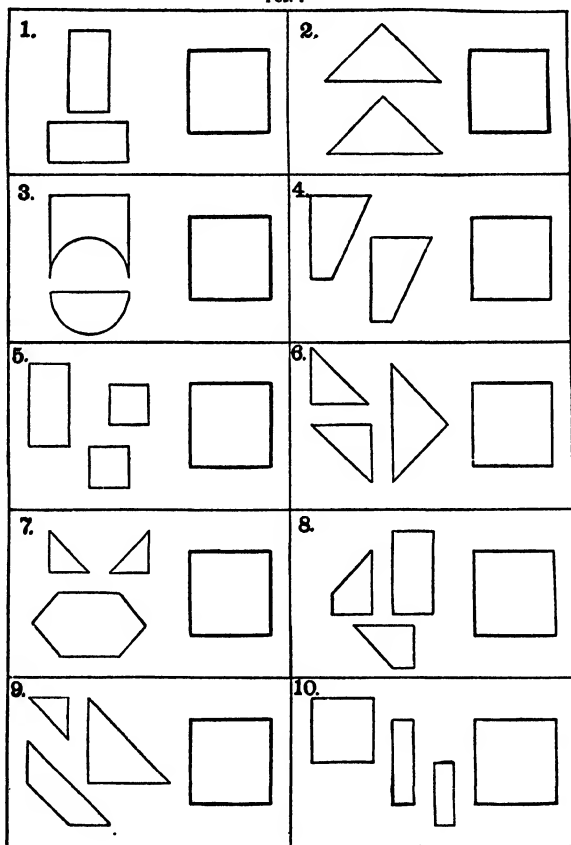
6.

Test 5

650 650	10243586 10243586
041 044	659012584 659021854
2579 2579	888172902 881872902
8281 8281	681027594 631027594
55190 55102	2499901354 2490001534
89190 89190	2261050310 2261639310
658049 650849	2911038227 2011038227
8293017 8290517	313877752 313377752
63015991 63019991	1012938567 1012938567
39007106 39007106	7166220988 7162220988
09931087 09931087	3177028440 3177682440
251004818 251004418	468672663 468672663
290036013 290030013	9104329003 9194329003
30015992 300155992	3484637120 3484637210
3910066482 391006482	8388172536 8381722536
8510273301 8510273301	3120106671 3120106671
263136996 263130006	7611348879 76111345879
451152903 451152903	26357239104 26357239104
8239016275 8239016725	8819002341 8819002341
582039144 582039144	6371018034 6371018034
61538529 61388529	38779762514 38779765214
211915883 219915883	39008126537 39008126637
670413822 670143822	75038100398 75038100398
17198591 17198591	41181900720 41181900726
86482991 86482991	6543920817 6548920871



Test 7



PART II

MEASUREMENT OF INTELLECTUAL QUALITIES

CHAPTER IV

The Measurement of Mental Deficiency

I. What Is Mental Deficiency?

MENTAL deficiency (or feeble-mindedness) is defined in terms of intelligence. It is the presence of a relatively small amount of whatever we understand to be intelligence or mental alertness. If intelligence is the ability to carry on abstract thinking, then mental deficiency is the relative lack of ability to engage in such a process; if intelligence is the capacity for learning new things, then mental deficiency is the relative absence of such a capacity; and if intelligence is a quality of general adaptability, then mental deficiency is a relative inability to adapt. Exactly where we draw the line that marks off the mentally deficient is arbitrary. The best person below the line must be almost, if not quite, as good as the lowest person above it. We might say that a person is mentally deficient if his learning capacity is so low that he cannot graduate from elementary school. However, we might arbitrarily, and with just as much logic and justification, say that a person is to be labeled "mentally deficient" only when he is incapable of learning to feed and dress himself. Where the line is drawn has been determined largely by convenience in dealing with the results.

Early definitions of mental deficiency or feeble-mindedness, before the advent of mental testing and psycho-

logical study of intelligence, were likely to be legal or sociological. The *legal definitions* usually emphasized the lack of responsibility of the deficient one before the law. Sometimes they concerned the obligations of mill-owners and other industrial managers in selecting and training child workers in their plants. *Sociological definitions* of mental deficiency regard the ability of the individual to get along with his fellows, to manage his own affairs, to support himself in the world. The mentally deficient find it difficult to adjust themselves in the social world; they are unable to earn a living for themselves; they become involved in crimes; their behavior is often anti-social; and they may become a burden for public support and maintenance. *Medical definitions* for mental deficiency emphasize the neurological basis of the lower capacity or ability. Tredgold's definition illustrates this type: "We may accordingly define amentia as a state of restricted potentiality for, or arrest of, cerebral development, in consequence of which the person affected is incapable at maturity of so adapting himself to his environment or to the requirements of the community as to maintain existence independently of external support."¹

With the psychologists' interests, many *behavioristic definitions* of mental deficiency have arisen. These are based upon a comparison of the behavior of the mentally deficient with that of the normal individual. They are definitions in terms of how he does things as compared with others of his own chronological age. For such purposes comparisons are made of educational attainments and of performances on various mental test problems. Psychological studies have also been influential in giving us a definition of mental deficiency in terms of

¹ Tredgold, A. F., *Mental Deficiency* (fifth ed.), Wm. Wood and Co., New York, 1929.

a *mathematical, statistical, or percentage* concept. Such definitions are the direct result of mental testing and of the hypothesis of a normal distribution of intelligence in the whole population. Mental deficiency is regarded as the lower end of the distribution curve, and is defined accordingly as a certain percentage at the lower end. The exact percentage is arbitrary. If taken low enough to mean cases commonly found under institutional care, it would amount to less than one per cent of the distribution of all individuals. If taken to include all those incapable of supporting themselves without assistance, it would probably be two or three per cent. Common practice among mental test experts is to draw the line so as to class about one per cent in the group designated as feeble-minded.

So there are several viewpoints from which we may consider the quality of intelligence, and each viewpoint gives us a somewhat different definition. All the definitions are important in giving us a total view of the problem of mental deficiency. Measurement in psychology concerns itself chiefly with the psychological and statistical definitions, but the application of the measurements cannot neglect legal and sociological aspects of the problems.

II. Degrees and Types of Mental Deficiency

On the basis of psychological measurements, mental defectives are often classified into three groups according to degree: (1) the *moron*, (2) the *imbecile*, and (3) the *idiot*. The moron is one who at maturity shows a mental development only to the point of the normal 8- to 10-, 11-, or 12-year-old. (There is some disagreement among investigators as to the upper limit.) Imbeciles show a development to between the 3- and 7-year level; and idiots do not develop beyond a 2-year level in ability.

62 Measurement of Mental Deficiency

These classifications can be defined in terms of school or social capacities, but since we can measure mental ages by psychological tests, the age definitions are more useful.

Mental deficiency can also be classified into types according to causal factors. The *cretin* is a mental defective from thyroid gland deficiency, the *microcephalic* from having a small brain, the *paretic* from syphilitic infection of the brain, the *hydrocephalic* from fluid in the cranial cavity, and so on. From the standpoint of mental measurements these types are of little significance, since the measurement of mental ability is a quantitative estimation without reference to cause, and since these types are not homogeneous within themselves as to mental ability. A cretin may range from borderline dullness to idiocy; a paretic may be just below normal or a total idiot. Some of the types occasionally show normal or superior intelligence.

III. Why Measure Mental Deficiency?

Mental deficiency is a handicap to the individual in any situation in which he must adapt himself to new surroundings, learn new modes of behaving, adjust himself to the behavior of others, or solve problems requiring more than the minimum of judgment or reasoning. Problems of mental deficiency may arise in the management of a child in the home; in the instruction of a child in school; in the disposition of a delinquent in a juvenile court; in the guidance of an adolescent toward a vocational career; in the handling of an inmate of an institution for the mentally disordered; or in the selection and placement of employees in an industrial plant. In all of these instances and in many others the objective measurement of the mental deficiency may be very useful.

Application of a definite measuring stick to the mental abilities of the problem individual may point out deficiencies not evident through general observation; and even though the presence of deficiency may be generally recognized, a measurement of the degree of defect is an important guide in the best solution of the problem. In school, the definite measurement of the amount of a child's deficiency may indicate whether he will profit more from instruction in the regular grade of a school or from instruction in a special atypical school. In an institution, measurement of amount of deficiency of a patient may suggest beneficial means of therapy and occupational training. In industry, a knowledge of the level of mental capacity of a failing employee may tell the personnel officer what shift in duties is necessary if the employee is to continue in the organization with a reasonable expectation of success and efficiency. Without measurement, mental deficiencies often exist and cause various problems in our relations with human beings without our suspecting that the deficiency is at the bottom of our trouble; we may confuse it with another defect or deficiency, or it may be covered or hidden by some other trait. Poor vision or poor hearing may simulate feeble-mindedness in a school child; a glib tongue and a pleasing manner may conceal a deficiency in an adult.

IV. Types of Tests Suitable for Measuring Mental Deficiency

The nature of mental deficiency suggests the following guides in the selection of mental tests to be applied to those low in the scale of intelligence:

1. *The tests should not require too great concentration and sustained attention on the part of the subject.* Men-

tal defectives are likely to give poor attention to the problems presented to them. If left to their own inclinations to answer a list of questions of a mental test, they will often fail because of inattention rather than actual inability to arrive at correct answers. Sustained attention may be a correlate of intelligence, but attention probably should not be the sole determining factor in arriving at an estimate of the mental level of a subject. This requisite of tests for measuring the mentally deficient favors the use of *individual* tests as compared with group tests. If group testing is done, only small groups should be tested at one time.

2. *Tests should not depend too much upon schooling.* The mentally deficient are likely to be backward in schooling. A mental test in terms of arithmetic, vocabulary, and reading, which may be suitable for superior persons with equal backgrounds of education, is likely to be primarily an educational achievement test for the mentally deficient. In view of this difficulty it is often desirable to choose a mental test which is given orally, and therefore does not depend upon the ability of the subject to read and write.

3. *Tests for the mentally deficient should not depend too greatly upon language ability.* General intelligence as we ordinarily define it does not manifest itself alone in language. It also manifests itself in dealing with more concrete things, and the deficient individual is perhaps more likely to be able to express himself through a concrete medium. Linguistic expression also has the disadvantage of being too closely related to formal schooling. We thus find the mentally deficient often being tested by picture tests, design tests, or manual-performance tests.

Following are some samples of representative tests which have been found useful in measuring the mentally deficient. These generally include (1) the individual age scales; (2) performance tests; (3) picture and design tests; and (4) simple written group tests.

1. THE INDIVIDUAL AGE SCALES

It seems appropriate that we examine first that type of measuring instrument which has proved most valuable in measuring mental deficiency. This instrument is exemplified in the Binet-Simon Scale and its various translations and modifications—all individually and orally administered scales.

We have already become acquainted with Binet because of his importance in the history of psychological measurement. It is of particular interest to us now that Binet's scales for measuring intelligence were developed primarily to solve certain problems which he faced in dealing with mentally deficient school children; and that his original scales were particularly suited to measuring such children. Indeed, Terman, in his American revision of the test, found it desirable to extend and revise the upper end of the original scale before it was suitable for testing at superior levels of ability.

In many ways, Binet's efforts in developing his intelligence scale represented a marked departure from previous and contemporary efforts at testing mental traits. He believed, *first*, that mental ability or intelligence should be tested in terms of complex mental processes instead of in terms of simple sensori-motor performance; *second*, that mental ability could be adequately measured only by presenting to the subject a variety of different tasks or problems; *third*, that a general mental ability rating

66 Measurement of Mental Deficiency

could be obtained by a composite of the subject's performance on a variety of tasks of varying difficulty; *fourth*, that the tasks or problems should depend on general rather than specific school experience; and *fifth*, that tasks and problems to be used in mental scales should be studied and standardized as to difficulty by trials on groups of individuals chosen at random.

Following these principles, Binet had by 1905 worked out and published his first scale. It consisted of thirty tasks or tests arranged in order of difficulty, as follows:

1. Ability to follow with the eyes a lighted match slowly moved before the eyes.
2. Prehension provoked by a tactile stimulus. (A piece of wood is brought into contact with the palm or back of the child's hand to see if he will seize it and carry it to his mouth.)
3. Prehension provoked by a visual stimulus. (Cube of wood merely shown.)
4. Recognition of food. (A piece of chocolate and a piece of wood are presented.)
5. Quest of food complicated by a slight mechanical difficulty. (Candy wrapped up in paper is presented.)
6. Execution of simple commands and imitation of simple gestures.
7. Verbal knowledge of objects. (Naming of parts of the body and familiar objects.)
8. Verbal knowledge of pictures.
9. Naming of designated objects. (Common objects on a picture must be named.)
10. Comparison of lengths of two lines.
11. Repetition of three digits.
12. Comparison of two weights. (3 and 12 grams.)
13. Suggestibility. (Asking for objects that are not present, etc.)
14. Definition of objects.
15. Repetition of sentences.
16. Comparison of two objects. (Giving differences between a fly and a butterfly, etc.)

17. Memory for things in a picture. (Picture shown for 30 seconds, after which child names objects seen.)
18. Drawing a design from memory. (Designs are shown for 10 seconds.)
19. Repetition of digits.
20. Resemblance of known objects. (In what way are a poppy and blood alike? etc.)
21. Comparison of lengths of lines.
22. Comparison of weights.
23. Memory for weights. (After blocks have been correctly placed in order of weight, one is taken away and the subject must find where the gap is.)
24. Rhymes. (Finding rhymes to a given word.)
25. Completion of sentences. (Supplying missing word to complete a sentence.)
26. Making up a sentence to contain three given words.
27. Comprehension of questions. (Twenty-five questions of varying difficulty to be answered.)
28. Reversal of the hands of the clock. (To be done from memory.)
29. Paper cutting. (Paper folded twice and triangular piece cut out. Subject must draw result without seeing unfolded.)
30. Definitions of abstract words.

This scale was not standardized according to age, but Binet gave rough standards of performance to be expected from normal children of various ages.

His own experience with the scale, and the criticisms of coworkers, led Binet to make two revisions—one in 1908 and one in 1911. These both constituted an important advance in having tests grouped according to age so that mental ability could be worked out as a definite measure expressed as "mental age." Their characteristics are essentially those of the Stanford Revision.

The Stanford Revision of the Binet Test. This revision of the Binet Scale stands, for several reasons, first among a number of American revisions: it is based upon the most thorough and extensive preliminary study of

68 Measurement of Mental Deficiency

the test material; it is well standardized as to both procedure and placement of the problems; it is the most commonly used of all the revisions. Terman, who directed the making of this revision, began his work around 1910, and in 1916 published it with an elaborate guide book² for use in administering and rating the test. This scale, like the original Binet Scale, has the tests or problems grouped according to age at which they can be normally answered; it also is rated in terms of a raw score of mental age. Terman made the calculation of mental age somewhat easier by the addition of tests enough to make six tests for each age level (with some exceptions in the upper ages). A summary of the tests at three of the ages in the Stanford Revision follows:³

YEAR III

1. Ability to point to parts of the body.
2. Ability to name familiar objects.
3. Enumeration of objects in pictures.
4. Ability to give sex.
5. Ability to give last name.
6. Ability to repeat 6 to 7 syllables.

YEAR VI

1. Ability to distinguish right and left.
2. Indication of missing parts from pictures.
3. Ability to count thirteen pennies.
4. Comprehension. Ability to answer questions.
5. Ability to name coins.
6. Ability to repeat 16 to 18 syllables.

YEAR XII

1. Vocabulary. Ability to define 40 of 100 test words.
2. Definition of abstract words.

² Terman, L. M., *The Measurement of Intelligence*, Houghton Mifflin Co., Boston, 1916.

³ Quoted by permission of Houghton Mifflin Company.

3. Superior plan for finding lost ball in circular field drawn on paper.
4. Ability to straighten out dissected sentences.
5. Interpretation of fables.
6. Ability to repeat five digits backwards.
7. Interpretation of pictures.
8. Ability to give similarities between three words.

The calculation of mental age. In the Stanford Revision of the Binet-Simon Test (up to age X—the most used ages) there are six tests per year, so that each test passed is credited as two months of mental age. Above the ten-year level fewer tests are available in the scale, so that each correct one is credited as three, four, five, or six months. In the actual rating of a test the subject is given credit for the last year which he gets entirely correct, with no errors in years below, and to this basal age is added the months' credit for additional tests answered above this. For example, if a child answers all questions correctly through year VI, answers five out of six correctly in year VII, three out of six in year VIII, and none above year VIII, his Basal Mental Age is six, and he receives ten months' credit in year VII and six months' credit in year VIII. His total mental age is, therefore, six years sixteen months, or seven years four months.

Average adults attain a rating on the Stanford Scale of XIV to XVI. Terman first gave XVI as the average adult level; later testings (particularly the Army testing during the World War) indicate that it is nearer XIV. Adult testing at the normal or superior level, however, is rarely done by a Binet type of test, so that the standards at the upper level probably have not been so well checked.

Intelligence quotient. Terman introduced with his revision of the Binet Test a new mental measure which

was designed to give more meaning to the mentality scores derived from testing. This new measure was the Intelligence Quotient (often abbreviated I. Q.). It is calculated by dividing the individual's mental age, as determined by testing, by his chronological age. The mental age of adults as indicated by the test is divided by the age which we take as representing the average adult. The I. Q. measure amplifies the significance of the mental age, since if we are without knowledge of the chronological age of the testee the same mental age may represent inferior, mediocre, or superior ability. For example, a mental age of 10 is superior in an eight-year-old; is average or normal in a ten-year-old; and is inferior in a twelve-year-old. As compared with mental age, the Intelligence Quotient is an expression of relative ability, signifying degree of "brightness" in a single measure. It remains fairly constant throughout life (since mental age and chronological age increase together up to the limit of mental growth or to "adult" age).

As we have already suggested, mental deficiency is often defined in terms of I. Q. level—and the I. Q. of 70 has been widely, almost universally, accepted as the arbitrary dividing line between normal (or at least only borderline) ability and definite deficiency or feeble-mindedness. If 70 is regarded as the dividing line for feeble-mindedness, the upper limit for feeble-mindedness in adults is a mental age of between 9 and 10. Terman worked out percentages of randomly selected individuals possessing mental ability at the various I. Q. levels. According to his tables, about one per cent of the population is deficient enough to fall into the feeble-minded grouping.

The following two examples from Terman's⁴ feeble-minded group illustrate testing at this low level:

⁴Terman, L. M., *op. cit.*, pp. 83 and 85.

R. H. Boy, age 14; mental age 8-4; I. Q. 60. Father Irish; mother Spanish. Family comfortable and home care average. Has attended school eight years and is unable to do fourth-grade work satisfactorily. Health excellent and attendance regular. Reads in fourth reader without expression and with little comprehension of what is read. Fair skill in number combinations. Writing and drawing very poor. Cannot use a ruler. Has no conception of an inch.

R. H. is described as high-tempered, irritable, lacking in physical activity, clumsy, and unsteady. Plays little. Just "stands around." Indifferent to praise or blame, has little sense of duty, plays underhand tricks. Is slow, absent-minded, easily confused in thought, never shows appreciation or interest. So apathetic that he does not hear commands. Voice droning. Speech poor in colloquial expressions.

Three years later, at age of 17, was in a special class attempting sixth grade work. Reported as doing "absolutely nothing" in that grade. Still sullen, indifferent, and slow in grasping directions, and lacking in play interests. No appreciation of anything, but has mastered such mechanical things as reading (calling the words) and the fundamentals in arithmetic.

In school work, moral traits, and out-of-school behavior R. H. shows himself to be a typical case of moron deficiency.

A. C. Boy, age 12; mental age 8-5; I. Q. 70. From Portuguese family of ten children. Has a feeble-minded brother. Parents in comfortable circumstances and respectable. A. C. has attended school regularly since he was 6 years old. Trying unsuccessfully to do the work of the fourth grade. Reads poorly in the third reader. Hesitates, repeats, miscalls words, and never gets the thought. Writes about like a first grade pupil. Cannot solve such simple problems as "How many marbles can you buy for ten cents if one marble costs five cents?" even when he has marbles and money in his hands. Described by teacher as "mentally slow and inert, inattentive, easily distracted, memory poor,

ideas vague and often absurd, does not appreciate stories, slow at comprehending commands." Is also described as "unruly, boisterous, disobedient, stubborn, and lacking sense of propriety. Tattles."

Three years later, at age of 15, was in a special class and was little if any improved. He had, however, learned the mechanics of reading and had mastered the number combinations. Deficiencies described as "of wide range." Conduct, however, had improved. Was "working hard to get on."

A. C. must be considered definitely feeble-minded.

2. PERFORMANCE TESTS

The term *performance test* has generally been applied to the type of test in which the response is some motor or manual manipulation rather than a verbal response. Such tests do not depend upon language responses, and as a rule are so given that there is no necessity for dependence upon verbal directions as to procedure. When these tests are not so complex as to approach tests of mechanical aptitude, they are often useful for measuring general ability or intelligence. They are usually relatively simple as to level of difficulty, they practically never depend upon a medium of expression learned by formal schooling, and sometimes they test aspects of general ability unrevealed by the verbal tests. They have been widely used as supplements to the Binet Scale and other language tests in the measurement of mental deficiency. They have a usefulness also in testing various other special groups—such as the foreign-born who does not speak the language of the verbal tests, the deaf and blind, and the illiterate.

There are many types of performance tests. One of the most commonly used is the *form board test*, in which the subject being tested fits variously shaped pieces into

corresponding depressions or cut-out places in a board. This type of test is a very old device. Before the middle of the nineteenth century Seguin had devised a form board for use in training mentally deficient children, although the board was not utilized as a means of mental measurement until over fifty years later, by Norsworthy in the study of mentally deficient children. Another commonly used type of performance test is the *puzzle block test*, in which a number of variously cut pieces must be put together to form a picture or object—some-what on the order of our recently popular “jig saw” puzzles. Various other types of construction, object completion, and tapping tests have been utilized. Attainment on these tests usually depends upon accuracy of the manual performances and upon speed with which the acts are done, some of the tests being graded on first, second, and third trials.

There are available several performance scales for measuring intelligence. These consist of a combination of several different performance tests (usually ten or more), the group having been standardized so that total performance on the tests can be interpreted in terms of a mental age or some other intelligence measure.

The *Pintner-Paterson Performance Scale* is particularly worthy of mention because of its adequate standardization and wide use, especially with children. The scale is composed of fifteen tests, as listed below. See also Fig. 1.⁵

1. *Mare and Foal Board*. This is a picture board of a mare and foal with a number of cut-outs which the subject has to put in the correct places. It is very simple, resembling a child's game, and serves

⁵ Reproduced by permission of C. H. Stoelting Company, Chicago, Illinois.

as a very good introduction for children. Time and number of errors are recorded.

2. *Sequin Form Board*. Ten blocks representing common geometrical forms are to be placed in their appropriate places. The time of the shortest of three trials is recorded.
3. *Five-Figure Board*. Five geometrical figures, each divided into two or three pieces, are to be placed in

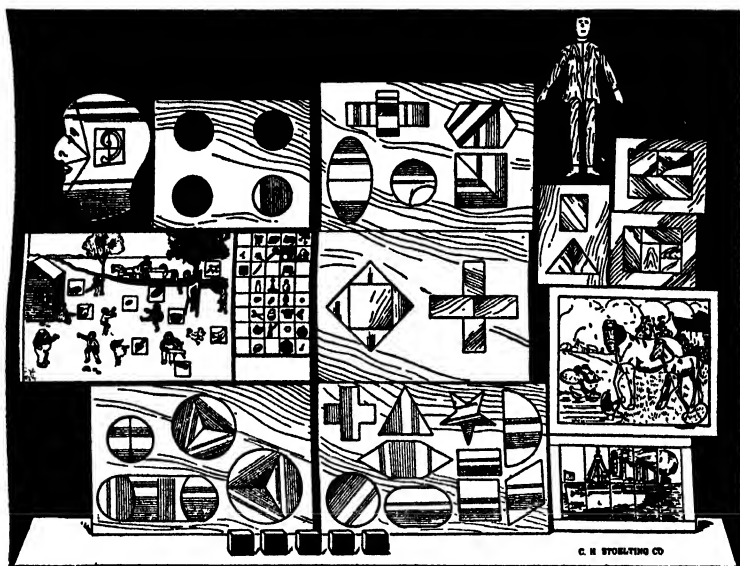


Fig. 1.—Tests in the Pintner-Paterson Performance Scale.

their appropriate places. Time and number of errors are recorded.

4. *Two-Figure Board*. Nine pieces to be placed in two spaces. Time and number of moves are recorded.
5. *Casuiet Board*. A more difficult board, consisting of four spaces into which have to be fitted twelve blocks. Time and number of errors are recorded.
6. *Triangle Test*. Four triangular pieces to be fitted into the board. Time and errors are recorded.

7. *Diagonal Test.* Five variously shaped pieces are to be fitted into a rectangular frame. Time and moves are recorded.
8. *Healy Puzzle A.* Five rectangular pieces are to be fitted into a rectangular frame. Time and moves are recorded.
9. *Manikin Test.* Subject has to put together legs, arms, head and body to form a man. There is no board into which the pieces fit. Quality of performance is scored.
10. *Feature Profile Test.* In the same manner as in the previous test, subject has to put together pieces to form a head. Time is recorded.
11. *Ship Test.* This consists of the picture of a ship cut into ten pieces of the same size and shape; these are to be fitted together properly into a rectangular frame. Quality of performance is scored.
12. *Picture Completion Test.* Subject is required to select the appropriate block out of many possible blocks to complete the picture. Quality of performance scored.
13. *Substitution Test.* A sheet of paper with rows of geometrical figures upon which the subject has to write the proper digit following the key at the top of the page. Time and errors are compounded into a score.
14. *Adaptation Board.* This is a simple test for measuring the ability of the subject to keep his attention upon a moving board. Number of correct moves is recorded.
15. *Cube Test.* Four cubes are tapped in a certain order, and the subject is required to watch and then imitate the movement. Number of combinations correctly imitated is recorded.

Each of the tests has been standardized individually so that a mental age is derived from the performance. The median mental age on the whole series is taken as the final mental rating.

76 Measurement of Mental Deficiency

Another well-known performance scale is the *Army Performance Scale*. This was worked out during the World War for testing foreigners, illiterates, and the recruits who were of such low ability that they failed the group tests. It consists of ten tests, three of which are also in the Pintner-Paterson Scale, as follows:

1. *The Ship Test*. (See 11 above.)
2. *Manikin and Feature Profile*. (See 9 and 10 above.)
3. *Cube Imitation*. (See 15 above.)
4. *Cube Construction*. Putting together small cubes painted on certain surfaces in such a way as to make a larger block painted on certain of its surfaces.
5. *Form Board*.
6. *Designs*. Copying a series of figures from memory.
7. *Digit-Symbol Test*. Same test as used in Army Beta Examination.
8. *The Maze*. Similar in principle to ones used in Army Beta Examination.
9. *Picture Arrangement*. A series of "Foxy Grandpa" pictures placed out of order to be arranged in story sequence.
10. *Picture Completion*. Missing parts of pictures to be supplied.

This Army scale was standardized in point scores, which were then translated into the Army letter rating scale. Equivalent mental ages for the point scores have been calculated. This scale has not been standardized on children, but the mental age equivalents are Binet mental ages.

How shall we evaluate performance tests as measures of mental capacity? They quite obviously measure mental ability in terms somewhat different from those in which the more familiar verbal tests measure it. Are these terms so different that we should not accept the

performance ratings as indicative of mental ability? Generally high correlations have been found between performance scale ratings and Binet mental ages, at least for the lower levels of ability where the performance scales are most often used. Buford Johnson, working extensively with the Pintner-Paterson Scale, found a correlation of .83 between performance mental ages and Stanford-Binet mental ages. Performance ratings and Stanford-Binet mental ages correlate to .84, as shown in the Army studies on American-born troops. Such studies lead us to rely upon performance scales as fairly valid measures of mental ability, particularly useful in those situations in which, for any reason, there is a language handicap in the subject.

3. PICTURE AND DESIGN TESTS

Like performance tests, picture and design tests are devised to avoid the great dependence of mental measurement upon verbal understanding and response. As compared with performance tests they have the advantage of usually being simpler to administer, of requiring less complicated equipment for testing, and of being suitable for group administration. Fig. 2 illustrates a test of this type. For Army testing during the World War, a whole test (the Army Beta) of different parts was developed utilizing only pictures, designs, and symbols. Chapter III contains illustrations from this test.

Porteus, in connection with his work with the feeble-minded at Vineland, devised a complete mental scale, standardized according to ages from 3 to 14, in terms of a single type of design—the maze. There is a maze for each age. The simplest for age 3 consists of a diamond with sides of parallel lines about a quarter of an inch apart. The subject is tested by his ability to trace

78 Measurement of Mental Deficiency

around the diamond, keeping within the parallel lines. At the more advanced years the subject is required to thread his way without crossing lines through a typical maze with many blind alleys. Fig. 3 illustrates mazes from the test.⁶ Porteus computes mental age on this

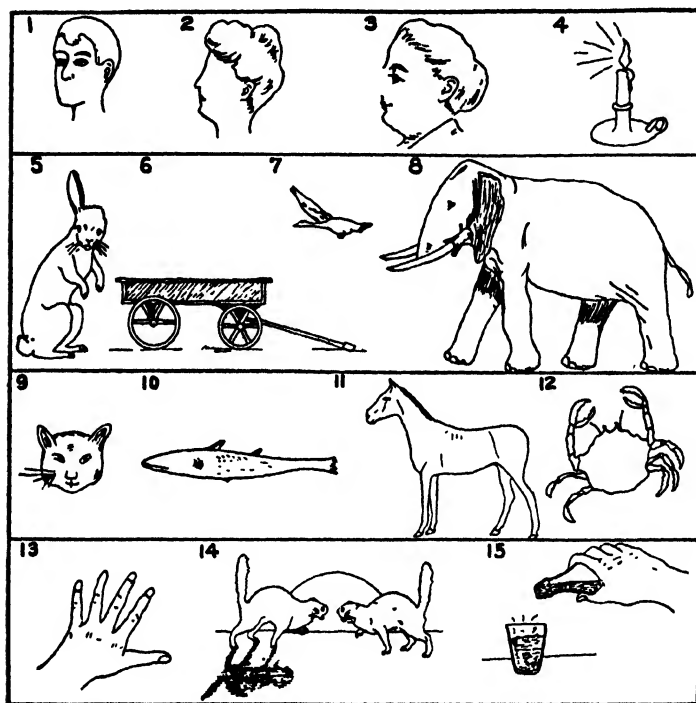


Fig. 2.—Test One—International Intelligence Test.

test by the highest test passed, with certain deductions for lower years that may be failed.

This scale is easily administered, usually attracts the

⁶ Reproduced by permission from *Maze Tests and Mental Differences*, the Vineland Training School, 1933.

interest of children and lower-grade mentality, and is applicable to groups who cannot be fairly measured by language tests. Its limitation as to type seems to be its chief drawback as a general intelligence test; it tests only one kind of response, while intelligence is manifested in many ways. However, in fairness to its author and users it should be stated that the test was designed as a sup-

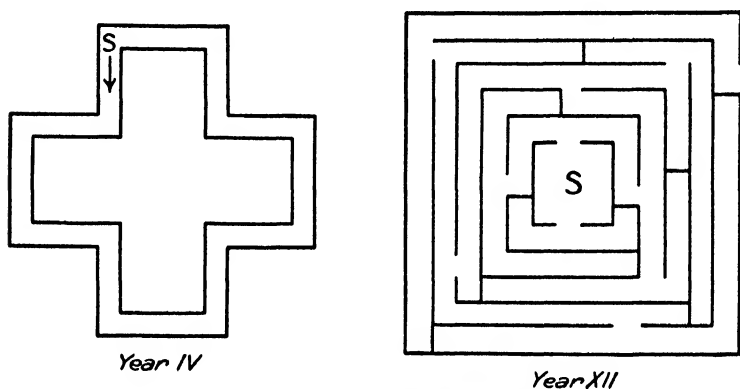


Fig. 3.—Porteus Mazes.

plement to such commonly used types as the Binet Scale. According to Porteus, the maze scale was devised to measure “planning capacity, prudence, and mental alertness in a new situation of a concrete nature.” Porteus believed the type of pre-considered action demanded by these tests was important in social as contrasted with educational efficiency.

The utilization of various picture and design tests in kindergarten mental measurements should be mentioned, although these are just as useful for measuring mental superiority in the very young as for measuring mental deficiency. The Haggerty Tests, the Pintner-Cunningham Test, and the Dearborn Tests are typical; several others, too, might be mentioned.

4. SIMPLE WRITTEN GROUP TESTS⁷

While the written group tests are not particularly adapted for measurement at the level of ability of the mentally deficient, the simpler ones and those which cover a wide enough range of ability to include simple items may sometimes be used. Such use, of course, always presupposes the ability to read and write in the subject being tested.

The Army Alpha Test, the written group test developed for testing recruits to the Army during the War, is a good example. This test measures intelligence over a wide range of ability—from markedly superior ability, possessed by those attaining high scores on the test, to definite feeble-mindedness, possessed by those attaining the scores in the lower range. The lower range of Army Alpha questions includes such easy ones as:⁸

How many are 20 boats and 9 boats? (Form 9)

If you save \$4 a month for 9 months, how much will you save? (Form 9)

If plants are dying for lack of rain, you should (water them, ask a florist's advice, put fertilizer around them). (Form 6)

A house is better than a tent, because (it costs more, it is more comfortable, it is made of wood). (Form 6)

The apple grows on a (shrub, vine, bush, tree). (Form 8)

The pitcher has an important place in (tennis, football, baseball, handball). (Form 9)

From Table II it may be inferred that any person obtaining a score as low as 21 on the Army Alpha Test

⁷ For a more detailed treatment of group mental tests, see Chapter V.

⁸ National Academy of Sciences, *Memoirs*, Vol. XV, pp. 228-234.

must possess a mental ability low enough to put him in the class of "mentally deficient" (I. Q. below 70, taking average adult mental age as 16). About one in ten of recruits tested by the Army Alpha made scores as low as 21. Some of these obtained a somewhat higher rating on individual tests.

TABLE II
EQUIVALENT MENTAL AGES AND I. Q.'S FOR
LOW ARMY ALPHA SCORES

<i>Army Alpha Score</i>	<i>Mental Age</i>	<i>I. Q. if Adult</i>
33	12-0	75
27	11-6	72
21	11-0	69
16	10-6	66
11	10-0	62
7	9-6	59
4	9-0	56
2	8-6	53

Many of the easy group tests commonly used in the elementary schools can be used also in measuring mental deficiency, provided the deficient subject is sufficiently literate to understand the test. For example, National Intelligence Test raw scores of below 70 represent definite deficiencies in those who have attained a chronological age of as much as 16.⁹

We might conclude that with proper safeguards as to ability of the subject to comprehend the procedure and the medium in which the test is expressed, easy group written tests may sometimes be used in measuring mental deficiency. We should be inclined, however, to give any person scoring low on such a test the benefit of an individual test.

⁹ See Table IV, page 98.

CHAPTER V

The Measurement of Superiority

MENTAL superiority—as genius—attracted the attention of many philosophers, historians, and poets of early times, and many such thinkers recorded their opinions as to the characteristics, sometimes even the psychological nature, of genius. Indeed it seems that the philosophers gave more attention to the genius than did our modern psychologists, scientific interest in high mental ability as a psychological problem having been late in developing as compared with the interest in deficiency.

[Carmichael,¹ in an address on the psychology of genius, admirably set forth these ancient views:

In folk lore and fable the genius plays a large part. In mythology he appears as a man set off from his fellows by a great chasm. He is a different species, not only quantitatively but in every sense qualitatively distinct. Usually his uniqueness has been considered as beginning in his peculiar origin. The Great Man was typically held to be the child of the supernatural, the offspring of god or demon.

This primitive notion—and who will gainsay Carlyle and the rest and claim that there are no reasons for its development—has persisted in modified form through the centuries. Certainly a supernatural theory of this sort was held by some of the greatest thinkers of an-

¹ Carmichael, Leonard, "The Psychology of Genius," *The Phi Kappa Phi Journal*, September 1934.

tiquity. Genius was essentially unique, a gift from another realm, in the view of Plato and Socrates. Similarly at all times the geniuses of religion, saints and prophets, have also been considered as supernatural; as men set apart by superhuman powers. In more recent times this transcendental view has not wanted supporters. Schopenhauer, for example, held that the genius was what might be called in biological terms a *psychological mutant*. The true genius, he asserted, perceived universals, whereas the common man perceived particulars and conceived universals—a view, incidentally, which seems to have nothing in its favor save that it was the view of the great Schopenhauer. Emerson, F. W. H. Myers, Hinkle, and more recently N. D. M. Hirsch have attempted by essentially mystical, or at any rate by non-scientific, arguments to develop a similar view of the true genius. Hirsch, for example, writing in 1931, claims, but without giving any evidence that would be likely to stand for a moment in the court of scientific psychology, that THE GENIUS—written in capital letters—has not common intelligence but intuition, not common volition, but a surging will to create, not ordinary emotion but ecstasy.

As opposed to this old but still vital *superior species* view of the superman is another ancient theory of genius which considers the superior individual to be still unique and unitary but abnormal. This is the view that genius is akin to madness. Thus Aristotle wrote, "Famous poets, artists, and statesmen suffer from melancholia or madness as did Ajax. In recent times such a disposition occurred in Socrates, Empedocles, Plato, and many others, but especially in our poets." Empedocles himself, strangely enough, also presented a similar view. Diderot proposed a "psychic rhythm" view of genius which held it to be essentially like manic-depressive insanity. Of all those holding this madness view of genius, Lombroso is probably most thoroughgoing and best known. He elaborately developed the theory that genius was a "biological type," a type ?
stigmatized in both mind and body and not unallied in

characteristics to other degenerate types; types, it now appears, on the basis of much evidence, that his imagination created rather than his science discovered. Others have held that genius is related to or caused by the physiological action of narcotics, alcohol, or the toxins of tuberculosis, and thus is comparable to the other mental abnormalities brought about by such drugs or disease. Havelock Ellis has guessed genius to be due to a sensitive nervous system and organic inaptitudes, such as muscular incoördination, which make social adjustment difficult. Freud contends that genius is akin to the neuroses and that it involves an atypical development of the sex life of the individual. Contrasted with this, Adler, another psychoanalyst, presents the view that genius is the result of an essentially pathological longing for superiority which leads the personality to flee from reality and become, through compensation, sometimes eminent, more often partly or fully insane. It is surely true that writers like Nisbet, Sanborn, and Kretschmer who have elaborately supported this madness position have found many cases to substantiate their theses. Alexander, Caesar, Napoleon, among the generals, many of the great philosophers, musicians, and painters were unstable personalities, epileptics, given to hallucinations or to hysterical outbursts. William Blake, to take but a single example, had innumerable visual hallucinations. Like Luther he really saw the devil. Once it is asserted that Blake sketched the ghost of a flea, invisible to others, complaining that its moving mouth interfered with his drawing. Swift, Johnson, Shelley, Byron, Tasso, all acted in a manner that today would lead their friends to say that they should have the care of a physician skilled in mental hygiene.

Today we do not look upon mental superiority and its highest manifestation, genius, in this light. We do not look upon the highly intelligent as endowed with any supernatural traits or any abilities the possession of

which cannot be explained by the same biological laws of inheritance that explain other ability levels. We do not look upon genius as related to madness or insanity. We recognize quirks of behavior or peculiarities in the genius as resting only upon the fact that because of his higher ability he is less likely to be interested in the things of the average intellect and for that reason may occasionally set himself apart as "peculiar." We look upon the genius or person of superior ability not as representing a qualitative difference in traits, but only as an individual differing quantitatively from his fellow beings; only as one possessing a higher degree or amount of the same types of abilities seen in the average or even in the mentally deficient.

Superiority might well be defined in those same terms that we used in defining deficiency. Sociologically, mental superiority connotes an exceptional capacity for getting along in the world—for adjusting oneself to one's surroundings—for success in general. Educationally, mental superiority manifests itself in high capacity for learning or high capacity for school success. Mathematically or statistically, mental superiority designates that certain percentage of individuals falling at the upper end of the scale as we measure intelligence by psychological tests. It is this last definition or view that has been the basis of recent studies of mental superiority.

I. Development of Interest in the Superior

1. **Early studies of special cases.** Before the advent of psychological tests, interest in and study of mental superiority were limited chiefly to cases of child "prodigies" who from time to time attracted enough attention so that their cases were recorded by interested educators. These children were usually studied by some educator

who became interested in seeing what could be accomplished by this or that method of education. One of the earliest records of a case of this type is contained in a German book published in 1779—*The Life, Doings, Travels, and Death of a Very Clever and Very Well-behaved Four-year-old Child, Christain Heinrich Heineken of Luebeck, Described by His Teacher, Christain von Schoeneich*. Among the remarkable achievements recorded for him are: the ability at ten months to name things in pictures; the ability at one year to recite stories from the books of Moses; and the ability at four years to read, add, subtract, multiply, and divide. He is credited with knowing French and German, much geography, and some 1500 Latin sayings. Unfortunately this prodigy died before the age of five and his accomplishments could not be followed to maturity. Among other infant prodigies written about before the days of mental testing were one Karl Witte, son of a German pastor; John Stuart Mill; and Lord Kelvin. It is difficult to estimate what mental test records might have shown for these young persons of superior mentality, but it seems certain that any quantitative rating would have been high.

Sir Francis Galton's study of the inheritance of genius deserves mention in any discussion or study of high ability. His book *Hereditary Genius*, published in 1869, is an account of his study of several hundred individuals of "genius" ability selected from among famous persons who had lived in Europe. His general purpose was to show, by a careful study of family background, that high ability was an inherited quality. Aside from the intrinsic value which his book has in relation to the question of inheritance, it has significance in the chronology of measurement of mental qualities in that it focused attention upon the existence of marked individual dif-

ferences in mental traits, and in that Galton suggested the feasibility of a scale for designating degrees of ability.

2. **Studies of childhood of great men.** More recently, systematic studies have been undertaken of the childhood of great men. These studies have been prompted by generally increased interest in the traits of genius and by efforts to refute the erroneous popular fallacy that geniuses often arise from stupid childhoods or have grown out of youthful lives of physical weakness, emotional peculiarity, and similar traits. The best of these studies is that of Cox.² From a study of boyhood records (largely from biographical sources) she attempted even to arrive at estimated I. Q.'s for her list of eminent men. She did not arrive at estimates of I. Q. for all of the eminent men that would place them in the level of "genius" intellect as we use the term in mental testing today, but her average I. Q. for the total 300 studied was 135, an I. Q. attained in less than one in a hundred tests today. While her data cannot be free from error (owing to the manner of their determination), they show rather clearly that the childhood of most eminent men of today and the past has been very similar in achievement to that which we now find in the child whose superiority has been adjudged by one or more mental tests.

3. **Development of interest in the superior school child.** Interest in the superior school child was much slower in development than interest in the subnormal child. This delay was due partly to the nature of the earlier mental tests, which were not adapted so well to the measurement of superior ability as to lower ability. The original Binet Test did not measure ability accurately or well at the upper levels (at least not for those

² Cox, C. M., "The Early Mental Traits of 300 Geniuses," *Genetic Studies of Genius*, Vol. II, Stanford University Press.

above a very young chronological age). The same criticism can be made of many of the earlier tests patterned after the Binet and of some of the earlier group mental tests. A second reason for late development of interest in the superior school child (undoubtedly the basic reason also for the types of tests developed in the early work) is the fact that he is not the school problem that the sub-normal child is. While he may get into mischief by lack of pursuits to test his full powers, the superior child does not draw attention because of lack of progress and inability to learn. It has been stated that the superior child has been discovered by the intelligence test, and this has certainly been true of many a superior child whose performance would not have been spectacular without the extra spurs to attainment that we can offer him through knowledge of his superiority.

Today the psychologist's and the educator's interest in the superior school child is attested by the various schemes of classification through which we attempt to vary instruction to fit the level of ability, by the enriched curricula offered the bright child, by the rapid progress allowed him through the grades, and by the various special studies of his mental, emotional, and physical make-up.

✓ 4. **Terman's studies of genius.** The important work accomplished by Terman in studying superior ability should not be left unmentioned. Pintner³ has given an excellent brief summary of this work.

About 1,000 children above I. Q. 130 were selected for study, and these are compared with children whose I. Q.'s were normal. It was found that among the

³ Pintner, Rudolph, *Intelligence Testing*, Henry Holt and Co., New York, p. 361.

gifted the ratio of boys to girls was higher than that in the general population. In racial origin these California children were found to be mainly of Western European and Jewish stock. The Jewish stock contributed about twice that expected from the total Jewish population of the areas investigated. The average social status of the families was much higher than that of the average family. In general the family incomes are fair, and they live in superior neighborhoods, but there are isolated cases from very poor families in inferior neighborhoods. These children come from families where there are distinguished relatives in much greater proportion than would be found in the average family. The vital statistics of the families show a healthier than average stock, with few cases of insanity or feeble-mindedness. The anthropometric measurements show the gifted group physically superior. The medical examinations show them also superior to average children. In school progress they are 14 per cent of their age above the norm in grade location, and 48 per cent of their age above the norm in intelligence, so that they are underpromoted to the extent of 34 per cent. Their school marks are better than those of ordinary children. The gifted are no more uneven in their school abilities than ordinary children. Their occupational ambitions are higher than those of the control group. In general they have the same type of interests as ordinary children. They make more collections, particularly of a scientific nature. Their play interests are in general like those of the control group, with a somewhat greater interest in plays that require thinking. They are mature in their play interests, showing a greater liking for quieter and less sociable games. These gifted children read a great deal more than does the average child. The average gifted child of 7 reads more books in two months than the average control child up to age 15, and the range of reading is much wider. In character and personality tests they are very superior, about 85 per cent of the gifted being above the median of the control group.

This work of Terman, briefly summarized above, is important because it gives us a rough picture of the gifted child and clears away the old ideas with reference to the sickly prodigy and the puny bookworm. What we note is the tendency for desirable traits to be positively correlated. Along with high intelligence goes general all-around superiority. Of course, there are individual exceptions. There are children of high I. Q. who are physically below normal, who do not like active games, who are nervous or vain and conceited just as there are normal or subnormal children with similar undesirable traits. But such undesirable traits do not tend to accompany superior intelligence. On the contrary they are less likely to be found among children of high I. Q.'s than among children of average or below-average I. Q.'s.

5. Binet Test applied to measurement of the superior child. The Binet Test and its revisions or modifications are often utilized in measuring the superior child, especially the preschool child or young child who has not yet learned to read enough to take a written group test. In fact for the majority of the special studies of superior children the Binet type of test has been used as the basis of selection of subjects because of its generally accepted validity and reliability.

The following are summaries of two interesting cases of superior children tested by Terman ⁴ with the Stanford Revision of the Binet Test:

E. B., Girl, age 7-9; mental age 10-1; I. Q. 130.
E. B. was selected by the teachers of a small California city as the brightest school child in the city (school population about 800). Her parents are said to be unusually intelligent. E. B. is in the third grade, a year

⁴ Terman, L. M., *The Measurement of Intelligence*, Houghton Mifflin Co., Boston, 1916, pp. 98 and 101.

advanced, but her mental level shows that she belongs in the fourth. The test was made as a demonstration test in the presence of about 150 teachers, all of whom were charmed by her delightful personality and keen responses. No trace of vanity or queerness of any kind. Health excellent. E. B. ought to be ready for high school at 12; she will really have the intelligence to do high school work by 11.

E. F., Russian boy, age 8-5; mental age 13; I. Q. approximately 155. Mother is a university student apparently of very superior intelligence. E. F. has a sister almost as remarkable as himself. E. F. is in the sixth grade and at the head of his class. Although about four grades advanced beyond his chronological age he is still one grade retarded. He could easily carry seventh grade work. In all probability E. F. could be made ready for college by the age of 12 years without injury to body or mind. His mother has taken the only sensible course; she has encouraged him without subjecting him to overstimulation.

E. F. was selected for the test as probably one of the brightest children in a city of a third of a million population. He may not be the brightest in that city, but he is one of the three or four most intelligent the writer has found after a good deal of searching. He is probably equaled by not more than one in several thousand unselected children. How impatiently one waits to see the fruit of such a budding genius!

II. The Superior School Child

1. **Occurrence and characteristics.** There are available today a great many studies surveying the intelligence of large numbers of school children. Let us examine some typical studies in order to gain an idea of the prevalence of superiority among school children as well as to gain an appreciation of the general distribution of intelligence throughout school populations.

During his standardization of the Stanford Revision of the Binet Test, Terman measured 905 unselected school children by this individual test. He obtained percentages of children at various I. Q. levels as follows:

<i>I. Q.</i>	<i>Percentage of Cases</i>
56-65	0.3
66-75	2.3
76-85	8.6
86-95	20.1
96-105	33.9
106-115	23.1
116-125	9.0
126-135	2.3
136-145	0.5

Pintner reports eight other studies of elementary school children by individual tests. These are shown with the Terman study in Fig. 4. From one study to another there are slight differences in distribution, depending upon such selective factors affecting intelligence of groups as locality of the schools tested, occupational

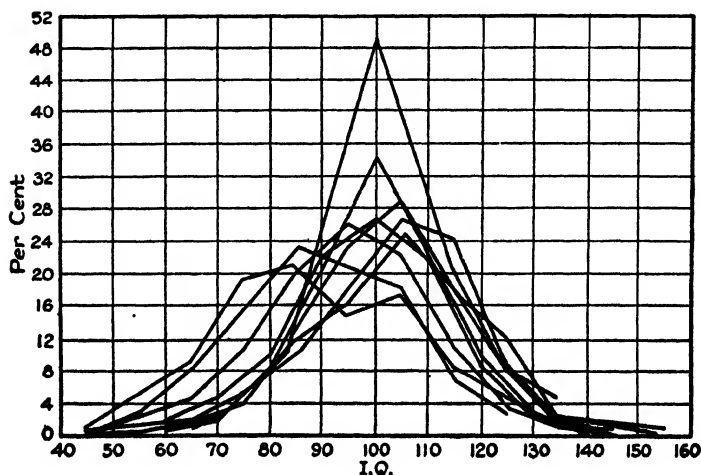


Fig. 4.—The Distribution of Intelligence. (Nine Studies.)

pursuits of the locality, etc. But the general characteristics of the various distributions are the same and the differences are rather small.

Superiority among elementary school children as judged from these distributions amounts to about 20 to 30 per cent of the total number if we include all those with I. Q. above 110, though some of these are not very superior; or to about one or two per cent if we include only those with I. Q.'s above 130 (usually called the "very superior" group).

Group testing of elementary school children shows results similar to those from individual tests. A report on over a thousand National Intelligence Tests in a New York public school shows 20.5 per cent with I. Q. of 110 or above; and 2.7 per cent with I. Q. of 130 or above.

What do tests in high school show? Surveys similar to those just discussed for elementary schools show an average I. Q. some 5 to 10 points higher, with an appreciable increase in the percentage of the total number of pupils who fall in the superior grades of intelligence. This is accounted for by certain factors at work which tend to "select" the pupils who go to high school. Practically all children start elementary school, but as we examine farther and farther up the line in school grade, we find that a larger and larger number of pupils have dropped out. In relation to mental measurements it is important to recognize that the selection process is partly on the basis of intelligence. The more intelligent tend to go to high school because they are capable of doing the work and have the requisite interest and ambition, while the less intelligent tend to drop out. Thus, with those at the lower intelligence levels tending to drop out, we find that the distributions for high school show greater amounts of superiority as compared with distribu-

tions of mental measurements of elementary school children. The contrast is even greater when we compare college distributions of mental measurements with public school distributions. Fig. 5 shows graphically the selective process of our educational systems as it affects the distribution of intelligence within the school group.

Superiority among elementary school children occurs about as frequently as it does in the general population; in high school about one and a half times to twice as frequently as in the general population; and in college probably three or four times as frequently. Aside from

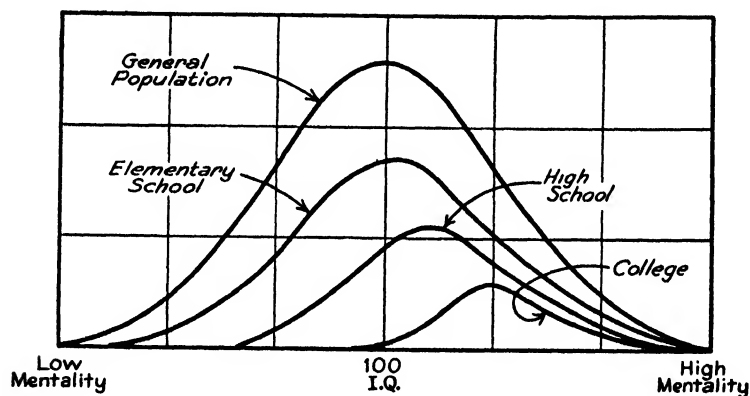


Fig. 5.—Diagrammatic Representation of the Selective Process in Determination of Intelligence of School Groups.

their test performances we find superiority in school pupils distinguished by good school work, though often not good enough to indicate the amount of superiority in ability; by grade placements often ahead of the pupil's chronological age; by above-average participation in extra-curricular activities; and by tendency to be found more frequently in schools of better communities and of cities rather than rural districts.

III. Tests Suitable for Testing School Groups

Superiority among school pupils is usually discovered by those mental tests which have proved most suitable for testing of school intelligence in general. Above the primary grades, where lack of reading proficiency is a handicap, these are generally group written tests. Those to be described as typical have been selected only because they are typical, not because they are distinctly better than many others which might have been selected.

1. The National Intelligence Test. This is one of the tests widely used for measuring mental ability in the elementary school. Its development was mentioned in Chapter III. The test covers 196 score points, divided approximately equally among five parts. The nature of the test is indicated by the directions and first few practice items in one of the forms.⁵

EXERCISE 1.—ARITHMETICAL REASONING

Find all the answers as quickly as you can. Write the answers on the dotted lines. Use the sides or bottom of the page to figure on.

1. How many cents are six cents and five cents? Answer-----
2. A girl earned 75 cents and spent 43 cents. How much did she have left? Answer-----
3. How many nickels make a dollar? . Answer-----
4. How many square inches are there in a card 7 inches long by 6 inches wide? Answer-----

⁵ From *National Intelligence Test, Scale A, Form 1*. Copyright by World Book Company, Yonkers-on-Hudson, New York. Reprinted by written permission of the publishers.

Measurement of Superiority

EXERCISE 2.—SENTENCE COMPLETION

Write on each dotted line one word to make the sentence sound sensible and right.

1. The apple is_____
2. Fish swim_____the water.
3. Boys_____girls like to_____ball.

EXERCISE 3.—LOGICAL SELECTION

SAMPLES	man	(body cane head shoes teeth)
	dog	(blanket chain collar <u>legs</u> <u>nose</u>)
	house	(cellar paint <u>room</u> servants <u>walls</u>)

In each row draw a line under each of the two words that tell what the thing always has.

1. table (books cloth dishes legs top)
2. apple (basket redness seeds skin sweetness)
3. shoe (button foot sole toe tongue)

EXERCISE 4.—SAME-OPPOSITE

SAMPLES	cold_____	<i>D</i> _____	hot
	big_____	<i>S</i> _____	large
	best_____	<i>D</i> _____	worst

If the two words mean the same, write *S* on the dotted line between them. If they are as different as can be, write *D* between them.

1. yes_____no
2. son_____daughter
3. light_____bright

EXERCISE 5.—SYMBOL-DIGIT

Make under each drawing the number you find under that drawing in the key. Do each one as you come to it.

KEY	Δ	\square	∞	+	Σ	P	G	H	\diamond
	1	2	3	4	5	6	7	8	9

Begin here:

∞	Δ	+	P	\square	G	+	Δ	Σ

The National Intelligence Test is available in four forms. It is well standardized as to procedure and results. Grade norms are given for the various forms; also age standards which make it possible to translate the point scores into mental age and I. Q. records. These are illustrated in Tables III and IV.⁶

TABLE III
GRADE NORMS (AVERAGE) FOR SCALE A,
NATIONAL INTELLIGENCE TEST

Grade	Number of Cases	Scale A
Low 3	2319	36
High 3	2008	49
Low 4	4608	63
High 4	2398	69
Low 5	5376	84
High 5	2154	89
Low 6	4592	102
High 6	2533	105
Low 7	4350	117
High 7	2132	124
Low 8	2832	130
High 8	1767	133

37069

2. **The International Intelligence Test.** This test, one more recently developed for use in certain racial studies,

⁶ *National Intelligence Test Manual of Directions*, World Book Co., Yonkers-on-Hudson, 1924, pp. 37 and 39.

TABLE IV
MENTAL AGE EQUIVALENTS FOR NATIONAL
INTELLIGENCE SCORES

<i>Score Scale A</i>	<i>Mental Age</i>		<i>Score Scale A</i>	<i>Mental Age</i>	
	<i>Yrs.</i>	<i>Mos.</i>		<i>Yrs.</i>	<i>Mos.</i>
46	8	6	92	11	5
47	8	7	93	11	6
48	8	7	94	11	6
49	8	8	95	11	7
50	8	9	96	11	8
51	8	10	97	11	9
52	8	11	98	11	10
53	8	11	99	11	10
54	9	0	100	11	11
55	9	1	101	12	0
56	9	2	102	12	0
57	9	3	103	12	1
58	9	4	104	12	2
59	9	4	105	12	2
60	9	5	106	12	3
61	9	6	107	12	4
62	9	7	108	12	5
63	9	8	109	12	6
64	9	9	110	12	7
65	9	10	111	12	7
66	9	10	112	12	8
67	9	11	113	12	9
68	9	11	114	12	10
69	10	0	115	12	11
70	10	1	116	13	0
71	10	2	117	13	0
72	10	3	118	13	1
73	10	4	119	13	2
74	10	5	120	13	3
75	10	6	121	13	4
76	10	6	122	13	4
77	10	7	123	13	5
78	10	8	124	13	6
79	10	8	125	13	7
80	10	9	126	13	9
81	10	10	127	13	10
82	10	11	128	13	11
83	10	11	129	14	0
84	11	0	130	14	1
85	11	0	131	14	2
86	11	1	132	14	4
87	11	2	133	14	6
88	11	2	134	14	7
89	11	3	135	15	1
90	11	4	136	15	6
91	11	5			

It consists of 200 points divided among eight parts. The following quotations from the test outline the parts: ⁷

TEST 1—RECOGNITION OF MISSING PARTS

Directions: In each of the pictures below some important part is missing. For example, the first picture has no mouth. Take each of the fifteen pictures and draw in the missing part. Don't take time to do beautiful drawing, but merely show that you know what part is missing. Work as fast as you can. (See page 78 for sample.)

TEST 2—INFORMATION AND OBSERVATION

Directions: If the statement is true, draw a circle around the *T*; if it is false, draw a circle around the *F*.

SAMPLE: *T* (F) The sun rises in the west.

T F 1. As a rule, the temperature is higher during the day than during the night.

T F 2. More thunder storms occur on hot days than on cold days.

TEST 3—MEANING OF WORDS

Directions: If a pair of words mean the same or nearly the same, draw a circle around the *S*; if they mean the opposite, draw a circle around the *O*.

SAMPLES: { little small (S) O
 good bad S (O)
1. high low S O
2. anger wrath S O

TEST 4—DISCRIMINATION

Directions: If the situation described below could be true or is possible, draw a circle around the *P*; if it could not be true or is impossible, draw a circle around the *I*.

SAMPLE: *P* (I) His younger brother was three years older than he.

P I 1. He could count before he could read.

⁷ Quoted by permission of the Center for Psychological Service, Washington, D. C.

- ## TEST 5—RELATIONSHIP

TEST 6—COMPREHENSION

Directions: Below are two sets of proverbs. For each proverb in Section B there is a proverb in Section A which means the same or nearly the same. On the line before each proverb in Section B write the number of the proverb in Section A which means most nearly the same.

[Proverbs follow]

TEST 7—REASONING

Directions: Solve the following problems. Place the answers on the lines at the right. Use the space at the bottom of the page for any figuring needed.

1. How many men are 6 men and 7 men?..... -----
2. How many toes has a man who has lost one
on each foot?..... -----

TEST 8—UNDERSTANDING DIRECTIONS

The directions tell you to indicate certain facts in the five columns at the right of the table. You will find all the information you will need in the table. Read each of the ten directions and do exactly what it tells you.

Table V shows norms for the International Intelligence Test.

TABLE V
NORMS FOR INTERNATIONAL INTELLIGENCE TEST

<i>Grade</i>	<i>Score</i>
Elementary School:	
Fourth Grade	39
Fifth Grade	55
Sixth Grade	76
Seventh Grade	90
Eighth Grade	101
High School:	
Freshman	119
Sophomore	129
Junior	142
Senior	148

3. The Otis Intelligence Scale. The Advanced Examination of Otis's Group Intelligence Scales is one of the most used mental tests for measurement at the high school level today. Its parts, indicated by name and sample questions, are outlined below.⁸ The test has been standardized for grade norms, age norms, and mental age equivalents.

TEST 1—FOLLOWING DIRECTIONS

SAMPLE PROBLEM: Write the fifth letter of the alphabet. ()

TEST 2—OPPOSITES

Directions: Look at the first word in each line, think what word means exactly the opposite of it, find that word among the five words in parentheses in that line and draw a line under it.

up (short, down, small, low, young)
hot (warm, ice, dark, cold, fire)

⁸ From *Otis Group Intelligence Test, Advanced Examination*. Copyright by World Book Company, Yonkers-on-Hudson, New York. Reprinted by written permission of the publishers.

TEST 3—DISARRANGED SENTENCES

Directions: The words on each line below make one sentence if put in order. If the sentence the words make is true, underline the word *true* at the side of the page. If the sentence they would make is false, underline the word *false*.

men money for work..... (true false)
 uphill rivers flow all..... (true false)
 ocean waves the has..... (true false)

TEST 4—PROVERBS

Directions: Read each proverb, find the statement that explains it, and put the number of that statement in the parenthesis before the proverb.

PROVERBS

- (3) The burnt child fears the fire.
 () Rome was not built in a day.
 () There is not smoke without fire.
 Etc.

STATEMENTS

1. Time is required to produce anything of value.
 2. Failure follows frequent change of plan.
 3. Unhappy experiences teach us to be careful.
 4. Those in disgrace always want to disgrace others.
 5. There is no result without a cause.
- Etc.

TEST 5—ARITHMETIC

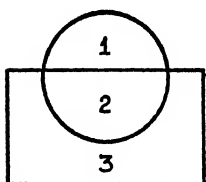
Directions: Place the answer to each problem in the parenthesis after the problem. Do any figuring you wish on the margin of the page.

1. If a boy had 10 cents and earned 5 cents,
 how much money did he have then? . . . () cents
2. If a man walks east from his home 7
 blocks and then walks west 4 blocks, how
 far is he from home? () blocks

3. If a wire 20 inches long is to be cut so that one piece is $\frac{2}{3}$ as long as the other piece, how long must the original piece be? . . . () inches

TEST 6—GEOMETRIC FIGURES

Directions: Each problem asks a question that is answered by a number. Write the answer to each problem in the parenthesis after the statement of the problem.



Look at the figure. What number is in the circle but not in the rectangle? ()

TEST 7—ANALOGIES

Directions: The first sample means: Finger is to hand as toe is to what? Underline the word on each line that should go in the parenthesis in place of the question mark.

finger: hand—toe: (?) foot, knee, arm, shoe, nail

TEST 8—SIMILARITIES TEST

Directions: Find the way in which the first three things on a line are alike. Then look at the five other things on the same line and draw a line under the one that is most like the first three.

hat, collar, glove . . . hand, cane, head, shoe, house
 rose, daisy, violet bush, red, plant, bed, pansy
 desk, bed, chair book, table, floor, pencil, coat

TEST 9—NARRATIVE COMPLETION

(The test consists of a story with words frequently deleted. These words may be filled in by choosing words from columns placed at the right of the story.)

Directions: For each numbered blank in the story, choose the best word of the three in the list having the same number as the blank. Underline the word you choose. You may write these words in the blank spaces if you wish, but only the underlining counts. Do nothing about the blanks that are not numbered.

THE REWARD OF KINDNESS

Underline words here:

Once upon a 1 there
was a 2 that lived in a
3 etc.

1. Time place man
2. man lion dog
3. street garden forest

TEST 10—MEMORY

(First a story is read to the class. Then the class answers from memory questions about the story.)

Directions: Read each question, and if the right answer according to the story is *yes*, draw a line under the word *yes*. If the right answer is *no*, draw a line under the word *no*. But if you do not know the right answer because the story didn't say, draw a line under the words *didn't say*.

Was the story about a king? . . . (yes no didn't say)

Was the king's daughter 16 years old? (yes no didn't say)

Was she ugly? . . . (yes no didn't say)

IV. Mental Tests for the College Level

College students represent the most highly selected school group with which the mental tester deals. While college students are not usually measured in terms of mental ages and Intelligence Quotients, checks on students who have been tested earlier in their school career and are entering college show only rare instances of college entrants with I. Q.'s below 100, and we can recall by way of comparison that the average elementary-school intelligence quotient is about 100.

Immediately following the development and use of the Army mental tests during the World War, the Army

Alpha Test was administered to many college students, and the records from its use afford us a means of judging the superiority of college students as compared with the general population as measured in the draft by the Army tests. Various of the studies which have been published give college average scores on the test from about 80 to 160. The central tendency for all colleges is between 130 and 140, total Army Alpha point score. The average, based upon over a million tests, for the drafted man was 63; only about one per cent reached a score of 135. The average score for the white officers in the Army was 135, the same as the average for college students in general. While these comparisons may be a little overdrawn, owing to the greater facility which the student has for test procedures because of recent academic experience, the differences are wide enough, even with liberal allowances, to indicate that the college group is an extremely selected group in the direction of superiority.

A few examples of college mental tests used at the present time will be discussed.

1. **The Mental Alertness Test.**⁹ The author has used this test to considerable advantage in testing college freshmen at the George Washington University. It consists of five parts covering 200 score points. The parts measure Vocabulary, General Information, Arithmetical Reasoning, Reading Comprehension, and Learning Ability. The parts are rather similar in nature, except for a higher level of difficulty of questions, to the parts found useful at the lower levels of ability (elementary and high school). This particular test contains a rather unique method of testing learning ability through the use of a preliminary study sheet, from which the student learns

⁹ Published by the Center for Psychological Service, Washington, D. C.

during the test period certain things to be asked about later. The average college freshman attains a score of 125 out of 200 points; only 10 per cent exceed a score of 165.

2. **The Thorndike Intelligence Test.**¹⁰ Immediately after the war Thorndike began experimentation with mental tests for college students at Columbia University. This work led to the development of the *Thorndike Intelligence Examination for High School Graduates*, now available in many forms and published in a new form for testing each year. This test is the most comprehensive of the college mental tests, consisting of three parts, and containing a total of several hundred points. It requires almost three hours to give, as compared with approximately an hour for most other tests. The test contains an unusually large proportion of material measuring reading-comprehension ability, an ability of undoubted importance to success in pursuit of college work, and an ability which has been demonstrated to be closely correlated with intelligence in the literate individual.

The Thorndike Test has been extensively used in college testing throughout the United States, and many studies of its value have been made. In many instances exceptionally high correlations have been obtained between the Thorndike test records and success of the students in college work. Wood,¹¹ in an extended study of the test at Columbia University, reports a correlation coefficient of .67 between the test and scholarship in college, as compared with a coefficient of only .26 between scholarship in college and high school marks.

¹⁰ Published by the Bureau of Publications, Teachers College, Columbia University, N. Y.

¹¹ Wood, Ben D., *Measurement in Higher Education*, World Book Co., Yonkers-on-Hudson, N. Y., 1923.

3. **The American Council Psychological Examination.**¹² The American Council on Education publishes, for college and high school, a mental test prepared under the direction of Thurstone, which is becoming widely used in colleges. It contains five parts, designated respectively as *Completion*, *Arithmetic*, *Artificial Language*, *Analogies*, and *Opposites*. Like the Thorndike Examination this test is published in a new edition each year. It is more easily given and rated than the Thorndike test.

V. Measuring the Superiority of Adults Outside of College

Certain problems in the measurement of superiority in adults outside of school and college make the development of special tests for this purpose desirable, although few such tests have been developed. The unsuitability of the tests designed for college use rests primarily upon three objections: (1) They are likely to have too much of an academic flavor, reflected in their general nature and subject matter. (2) They are likely to depend too largely upon high-school content material. (3) They often emphasize speed to too great an extent. The validity of these objections rests primarily upon the fact that adult groups outside of school and college are usually very heterogeneous with regard to nature of academic background and as to recent experience of an academic nature.

Some of the mental tests developed for selecting employees might illustrate special adult tests. The United States Civil Service Commission under the direction of O'Rourke has devised a series of mental tests at four

¹² Published by the American Council on Education, Washington, D. C.

levels of ability for testing adult applicants for federal positions. The highest level of these tests reaches the ability requisite for high-grade professional and technical positions.

CHAPTER VI

Intellectual Measurement of the Insane

THE need for definite quantitative measurement in the field of mental disorders, or insanity, cannot be too greatly emphasized. Those who examine for symptoms of mental disorder are too prone to rely upon general observation and to regard such general methods as sufficient and accurate enough. This attitude has been fostered somewhat by the lack of accurate measuring devices for insanity symptoms and characteristics, but also by the too generally accepted notion that common sense is all one needs to appreciate the manifestations of mental disorder.

The complete measurement of mental disorder includes measurement of intelligence, of other mental traits which may be of particular importance in certain types of disorder, of personality, and of physical or physiological variants which are the basis of the disorder. For each of these we should know the expected measurement for the various classes of mental disorder. Unfortunately, no such standards are generally available. In fact, quantitative measurement in relation to insanity is only in its experimental stage.

This chapter will be concerned principally with mental tests in insanity. Personality tests are discussed elsewhere, as are also physiological methods of measurement useful in diagnosing insanity.

I. General Mental Tests

Measurement of mental level is important to the student of mental disorders, because in many of the major psychoses there is an accompanying deterioration in intelligence. In this connection, lack of intelligence should not be confused with insanity. The two are by no means synonymous. An individual may lack intelligence without being insane, as in cases of pure feeble-mindedness; or an individual may be insane while still retaining his general intelligence in most respects, as in the case of many individuals suffering from paranoia (a disorder characterized by persecutory delusions). Low intelligence simply implies inability to cope with abstract ideas, and inability to reason, make judgments, think logically, and generally adjust oneself to new problems. Insanity implies a condition involving pronounced changes in aspects of the individual's personality. In connection with mental disorders, a knowledge of changes in intelligence may be of considerable diagnostic and prognostic value. Marked intellectual deterioration usually means considerable progress in the disease and a rather insidious type of disease. Prognosis for recovery is markedly diminished if the disease is accompanied by deterioration of intelligence.

The most commonly used test for determining the intelligence level of the psychotic patient is the Binet Test—in America, the Stanford Revision of this test. Next to this test come the various performance scales. The popularity of the Binet Scale for testing among the mentally disordered is related primarily to the fact that it is an individually administered scale. Of the scales for individual intelligence testing, it is probably the best test now generally available. For the younger years it

is very well standardized, and the test situations are quite well adapted to the general interest and level of the individual. There is, however, serious need for a scale better adapted to adult level. Even for the adult whose intellectual level is only seven or eight years, the subject matter of the test is unsuited to the person's general experience. For example, at year eight, two of the judgment questions are: "What's the thing for you to do when you are on your way to school and notice that you are in danger of being late?" and "What's the thing for you to do if a playmate hits you without meaning to do it?" The problems of making change in money, and many of the sentence comprehension tests, are also in child terms. The scale as a whole was constructed for children, and is largely in terms of child and school subject matter. In the absence of better adapted scales, it has been widely applied in testing adults who may have the intelligence of the child but live in a different way and deal with a set of problems entirely different from those of the child.

Practically the only intelligence scales better adapted in subject matter to adult problems are group scales entirely of the pencil-and-paper type. Such tests, however, demand too continuous a voluntary attention on the part of the individual, whose mental disorder may have seriously interfered with his attention power; or they contain too few different types of mental response; or they handicap the individual whose lack or remoteness of school experience precludes facility in reading and writing. For uniform results in testing the mentally disordered it is practically necessary to have an individual test of the Binet type, in which various kinds of mental tests are included.

The following examples of mental testing of two de-

112 Intellectual Measurement of the Insane

mentia praecox patients illustrate two interesting features often noticed in mental testing of the insane:

A case of intelligence deterioration. C. G., age 22 at time of commitment to institution as a dementia praecox. Her history describes her as having had many mannerisms from the early 'teens, such as tendencies to touch her face and teeth while constantly and continuously moving the muscles of her face. During a period before her admission she had become very untidy in her appearance, claiming that it was wasteful to spend money on clothes. She had lost two jobs as a result of peculiar emotional outbursts. Intellectually her record showed good school work done through the elementary school and beginning of high school, an indication without definite test results of intellectual level to the extent of at least a mental age of 14 or 15. At the beginning of second-year high school work the history shows a definite slipping in ability to do school work, and the patient had dropped out of school before the end of the year. Actual test results showed, on hospital observation and testing two years before commitment, a mental age of 13.5 years. At the time of the commitment the mental age by the Binet Scale was 12 years 2 months, with a prognosis of likelihood of persistent deterioration.

This case is typical of the intelligence findings in the deteriorating psychoses, including the groups dementia praecox, paresis, and senile psychoses. The severer cases of epilepsy should also probably come in this deteriorating group. In these cases successive mental tests practically always show a decrease in mental ability.

A case of marked "scatter" in results. E. F., age 25, diagnosed as dementia praecox. Description of the case shows typical emotional and personality picture of the dementia praecox. A Binet Test performance at age 25 shows the following results:

Intellectual Measurement of the Insane 113

8-year level and below	able to perform all tests.
9-year level	5 tests out of 6 correct; misses test of sentences containing three words.
10-year level	4 tests out of 6 correct; misses absurdities test and reading and report.
11- and 12-year level	5 tests out of 8 correct.
13- and 14-year level	1 test out of 6 correct; solves arithmetic prob- lems.
16-year level	1 test out of 6 correct; repeats digits.

The total mental age on the tests is 11 years, 6 months.

This case is cited as an example of extreme "scatter" in test performance. A test result shows little scatter when the individual rather abruptly reaches the limit of his performance, that is, when his misses are confined to a range of a few years on the scale. A test result shows great scatter when the misses are spread through a range of several years on the scale; in the case just cited the range covers eight years. Scatter on a Binet Test is generally considered as of significance in diagnosis of psychotic cases. Psychotic cases usually show greater scatter than normal cases or cases of pure feeble-mindedness. The exact reason for this is not clear. It may be due to disturbance of attention factors entering into the test performance, but is more probably due to unevenness of deterioration in the various aspects of mental ability.

Measurements in special realms or of special factors of mental ability have often proved valuable in dealing with the mentally disordered. We shall discuss memory tests and association tests as two of the most important of these measuring devices.

II. Memory Tests

Memory tests in clinical testing are often more than simply supplementary to general intelligence tests. There are certain mental disorders which manifest themselves primarily in memory disturbances—the various amnesias. There are other mental disorders which manifest themselves in part in special disturbances of memory, as the senile's poor memory for recent happenings and relatively good memory for happenings of long ago. These and other reasons make it desirable to have systematic quantitative means of observing memory disturbances.

Wells¹ has described what appears to be a very adequate memory test—one which he has used at the Boston Psychopathic Hospital. This test includes the following parts:

1. Old personal memories.
2. Current events.
3. Memory of common school information.
4. Rote memory of alphabet and figures to 20.
5. Memory of substitution problem.
6. Repetition of sentences.
7. Repetition of figures forward.
8. Easy town-state knowledge.
9. Hard town-state knowledge.
10. Figures backward.
11. Naming objects exposed for short time from memory.
12. Recognition, among a larger group, of twelve cards previously exposed.

This test has been standardized so that standards of normality are known. Wells reports tests of 179 insane

¹ Wells, F. L., *Mental Tests in Clinical Practice*, World Book Co., Yonkers-on-Hudson, N. Y., 1927.

individuals in which averages of the group as a whole showed performances about equal to normal performances in parts 1, 2, 3, 4, 8, 11; performances somewhat below normal in 6, 7, 12; and performances considerably below normal in 5, 9, 10. When classified into groups of different types of insanity, the greatest departures from normal were shown by the "Organic Brain Condition Group" and "Paresis Group" (mental disorder from syphilitic infection).

III. Association Tests

Association tests have been used for indicating various mental tendencies. They have been useful in studying disturbances of flow of thought; for giving a general indication of mental capacity; and for discovering the manner or process of building up tendencies to reaction in some of the pseudo-mental disorders which are simply cases of bad habit formation. Considerable information on associations of which a person may be capable can be brought out in conversation with him, if the conversation is directed to cover a sufficiently varied and wide field and if the words and expressions emitted in conversation are carefully noted. For convenience, however, and for more accurate investigation, several kinds of association tests have been devised. The most useful arrangement for this purpose is a standardized list of words (usually 100) to each one of which the subject must respond with the first word that comes into his mind.

The Kent-Rosanoff list of words is one of the most used word association tests. Their list is reproduced in Table VI. Kent and Rosanoff have attempted to standardize their test for normality of response. Their standards were made up by giving the list of 100 stimulus

116 Intellectual Measurement of the Insane

TABLE VI

FREE ASSOCIATION TEST (KENT-ROSANOFF)²

1. Table	26. Wish	51. Stem	76. Bitter
2. Dark	27. River	52. Lamp	77. Hammer
3. Music	28. White	53. Dream	78. Thirsty
4. Sickness	29. Beautiful	54. Yellow	79. City
5. Man	30. Window	55. Bread	80. Square
6. Deep	31. Rough	56. Justice	81. Butter
7. Soft	32. Citizen	57. Boy	82. Doctor
8. Eating	33. Foot	58. Light	83. Loud
9. Mountain	34. Spider	59. Health	84. Thief
10. House	35. Needle	60. Bible	85. Lion
11. Black	36. Red	61. Memory	86. Joy
12. Mutton	37. Sleep	62. Sheep	87. Bed
13. Comfort	38. Anger	63. Bath	88. Heavy
14. Hand	39. Carpet	64. Cottage	89. Tobacco
15. Short	40. Girl	65. Swift	90. Baby
16. Fruit	41. High	66. Blue	91. Moon
17. Butterfly	42. Working	67. Hungry	92. Scissors
18. Smooth	43. Sour	68. Priest	93. Quiet
19. Command	44. Earth	69. Ocean	94. Green
20. Chair	45. Trouble	70. Head	95. Salt
21. Sweet	46. Soldier	71. Stove	96. Street
22. Whistle	47. Cabbage	72. Long	97. King
23. Woman	48. Hard	73. Religion	98. Cheese
24. Cold	49. Eagle	74. Whiskey	99. Blossom
25. Slow	50. Stomach	75. Child	100. Afraid

words to a thousand normal subjects and tabulating the responses of these subjects. Each normal word offered in response to a given stimulus word has a frequency index dependent upon the number of the thousand individuals giving that response. Responses not listed in the frequency tables are termed "individual" reactions. The standards afford a measurement of the tendency of an individual to respond along the same lines as his fellow-beings or to respond in ways peculiar to himself.

It is interesting to compare the "individual" reactions

² Reprinted, by permission, from *Manual of Psychiatry*, by Aaron Rosanoff, published by John Wiley and Sons, Inc., New York, 1927.

to the Kent-Rosanoff words of insane patients with reactions of normal subjects (see Table VII). Only about 7 per cent of the responses of normal adults are of the "individual" type, as compared with about 27 per cent of responses of the insane.

TABLE VII

RESPONSES OF VARIOUS GROUPS TO KENT-ROSANOFF ASSOCIATION TEST ³

<i>Subjects</i>	<i>% Individual Reactions</i>
1000 normal adults	6.8
247 insane adults	26.8
250 defective children ...	13.0
125 normal white children.	8.6

Murphy ⁴ has reported a study of Kent-Rosanoff responses of 250 normal individuals, 120 dementia praecox patients, and 82 manic-depressives. He classified their responses into thirteen categories, such as similarity, contrast, adjective-noun, rhyme and sound associations, etc. The manic-depressive gave more rhyme and sound associations than did the dementia praecox patients; but no marked association differences between the two insane groups are reported.

³ Ibid.,

⁴ Murphy, G., "Types of Word Association in Dementia Praecox, Manic-Depressive and Normal Persons," *American Journal of Psychiatry*, No. 2, p. 539, 1923.

CHAPTER VII

The Uses of Mental Tests

IN THE three preceding chapters we have discussed the development of mental tests and the nature of the tests which have been constructed. We shall now describe various uses made of these tests and indicate some of the results of their use in several fields. The investigations of the use of mental tests have been so numerous that it would be impossible to describe or even mention all of them. In our discussion we shall not attempt to do more than to select representative studies.

As we have already seen, the beginning of intelligence testing was related to the study of feeble-mindedness, particularly in the classroom. Mental tests later began to be used for dealing with mental deficiency met elsewhere—among delinquents or among workers. With the development of group testing methods and with the demonstration of the utility of mental tests for measurements at the level of normal and superior ability as well as of inferior ability, psychologists undertook studies of the mentality of school children in general, and began the application of mental tests to all levels of ability in the educational world. Still somewhat later came the application of mental tests in industry and business. It has lagged somewhat behind academic applications, but certain values of mental tests in the selection of employees and in the classification of workers have now been generally recognized; and mental measurements are

proving an important aid in the solution of problems confronting the personnel manager and the industrial executive. During much of this use of tests in schools, clinics, and industries, various special studies and research studies utilizing the tests have been carried on. Some of these have aimed at improving the tests themselves, and some have aimed at finding out the relationships between intelligence and various other factors.

I. Uses in College

College students have probably taken as many intelligence tests as any other group. College instructors have usually been the ones to develop the instruments of measurement, and for that reason they have been particularly interested in applying the tests to their own problems. Furthermore, the college student seems to have had more than his share of acting as the "guinea pig" for the trying out of new material. We thus find innumerable published studies discussing the uses of mental tests with college students.

It is difficult to make an accurate statement regarding the extent of use of mental tests in colleges. In 1923-24 Toops undertook an extensive survey in which he gathered information from colleges throughout the country regarding their use of mental tests. He then found that 60 per cent of colleges and universities in this country were making official use of mental tests. There seems to have been no similar survey at a more recent date, but it is likely that the figure at the present time would be somewhat over 60 per cent. It is also to be remembered that many of the colleges which make no official use of tests administer and use them informally at the behest of certain instructors, particularly in the psychology and education departments of the colleges.

There are probably very few colleges in this country at the present time in which mental tests are not administered and used in some way with a considerable number of the students.

The employment of mental tests for general and administrative purposes in college includes primarily four types of uses: (1) as a partial basis for admission; (2) in dealing with low scholarship cases in connection with decisions as to dismissals or probation; (3) in dealing with disciplinary cases; (4) in appraisal of transfer credits from little-known institutions.

The first of these uses is illustrated in a procedure which has been followed in the George Washington University. At this institution the student who has attended an accredited secondary institution and has attained a preparatory school record in the upper part of his class is admitted to the University on the basis of this record. However, the student who applies for admission from an unaccredited preparatory school or from a school whose standards are little known, or the student who applies from an accredited preparatory school but whose record is relatively low, is required to take a mental test and an educational achievement test as a basis for his admission. It happens quite frequently that by this method students of superior mentality and considerable chance of college success are discovered and admitted to college, whereas otherwise they might have been denied admission. An example of such a student may be quoted from a group of case studies made at this institution.

Case 19. Boy, age 18, graduated from a small high school with average grades. Interested in studying medicine. Applied for entrance to the University in September 1932. Percentile scores on the Thorndike

Intelligence Test and the Iowa High School Content Test administered on application were respectively 71 and 97. Admitted to college on the basis of these scores. Since he has been in college he has made a quality point index of plus 2.89 (almost a "B" average). He has worked and has paid a considerable proportion of his college expenses. He is a member of a social fraternity and of a professional chemistry fraternity. At the time of this study, which was at the end of his sophomore year, he was still registered in the University as a pre-medical student and was maintaining a good standing.

Mental tests are of considerable advantage in dealing with low scholarship cases in college. Low scholarship and low mental ability would suggest in most instances the advisability of dropping the student from the college or university and encouraging his pursuit of some other type of training. On the other hand, low scholarship and high mental ability might suggest the advisability of investigating and removing, if possible, other reasons for the low scholarship.

In addition to the four general and administrative uses, individual instructors and administrative officers in colleges often make many other uses of mental test results. During the year 1933-34 Remmers, of the Psychology Department of Purdue University, made a study of the uses which instructors made of their test results by circulating a questionnaire among the staff members of the University. Table VIII, quoted from Remmers' study,¹ shows the uses which the instructors indicated, with the percentage of the staff members making each of the uses.

Most of the studies of mental tests in colleges, and the

¹ Remmers, H. H., "Report on the Uses Made of the Freshman Entrance Test Results at Purdue University," *Bulletin of Purdue University*, Vol. XXXIV, No. 4, December 1933.

TABLE VIII

USES MADE OF MENTAL TESTS AMONG STAFF MEMBERS AT PURDUE UNIVERSITY

<i>Use</i>	<i>Per Cent Making Use</i>
Encouraging extra effort in the case of unmotivated bright students	46
Basis for individual consultation for improvement of study habits	45
Determining amount of academic work to carry; a minimum for high students, a maximum for low	30
Advice to sororities and fraternities <i>re</i> prospective pledges	25
For advising election of curricula or major subject	22
Encouraging capable students to undertake graduate work	20
Determining amount of work for self-support	19
As partial basis for dismissal for low scholarship . . .	14
Dealing with disciplinary (deportment) cases . . .	12
Making recommendations for scholarships or fellowships	12
Sectioning students according to capacity for progress	10
Appraisal of value of transfer credits	8
As partial basis for admission	6
Determining membership in honorary scholastic societies	5
Selection of assistants	5
Hiring student clerical help	5
Research purposes of graduate students or students in classes in "mental tests"	5
Appointment of students to non-academic offices .	5
Establishing control sections in research work . . .	4
Satisfy curiosity	2

type of study on which most of the uses are predicated, are those which investigate the relationship between intelligence test results and scholastic attainments of the students. Pintner² reports somewhat over seventy-five of the important studies that have been made of the relationship between mental tests and college grades. The

² Pintner, Rudolph, *Intelligence Testing*, Henry Holt and Company, New York, 1931, p. 302.

correlation coefficients between these two variables as obtained in the various studies show a central tendency of about plus .45. There are very few correlations below .30 and very few above .60. About the highest values that have been claimed for mental tests in predicting college grades are those announced by Wood in his investigation of the predictive value of the Thorndike Intelligence Test as utilized in Columbia University. He reports a correlation between the test and two-year scholarship of Columbia men of .67. This is markedly superior to the correlation which he reports for the same group between secondary school marks and college work, this correlation being plus .26. In general, we may state that the correlations between scholastic performance in college and mental test scores show that intelligence is one of the most important factors making for high marks. However, the relationships are sufficiently low to indicate that intelligence is by no means the only factor. Other factors, such as interest, health, effort, etc., also influence scholastic attainments.

II. Uses in Schools Below the College Level

In general, the lower schools in using mental tests have purposes similar to those of colleges. As in college, most of the uses are based upon the assumption that a measurement of mental ability is a good device for predicting attainments in school. Studies in high school and elementary school have shown that the relationships between grades in school and mental test records are about the same as relationships between scholastic attainment in college and mental test records. The correlations for these studies also are generally around .45.

The student of mental tests will notice certain uses and certain types of studies that are perhaps rather pe-

cular to the lower schools. One of the most important of these is the marked use that has been made of mental tests for classifying pupils into ability sections. Many of our large school systems today have established separate classes for the superior, the average, and the inferior child of a given grade. Usually these classifications are based either entirely upon intelligence testing or upon a combination of intelligence measurement and educational-achievement measurement. These classification schemes have proved to be of considerable value where variations in teaching method have been made to suit the mental ability of the group, or where separate curricula have been worked out for the different groups.

In the lower schools mental tests have frequently figured in special studies, as: survey studies in which whole school systems have been measured; studies which aim at making certain comparisons in mental ability of school children, such as comparisons from school to school or between city and rural or private and public schools; and studies which have aimed at finding out the amount of overlapping from grade to grade in the elementary or high school.

III. Uses of Mental Tests in Vocational Selection

Next to the use in schools, the greatest use of mental tests has probably been made in the selection of personnel in various industries and various departments of government. The use of mental tests for this purpose has lagged somewhat behind application in the schools, and even today we find the percentage of industries and personnel agencies in the government making use of mental tests considerably smaller than the percentage of schools making use of tests.

One of the most comprehensive studies of employment

methods used in American private personnel management was made by Stanley B. Mathewson.³ He sent questionnaires regarding personnel practices to 500 nationally known business concerns, picked at random. He received 195 answers. The 195 companies represented were located in 21 states and employed 2,391,000 workers, the numbers in the separate organizations ranging from 100 to 240,000. Forty-one different kinds of business enterprises were represented. In this survey he found that psychological tests were included among the employment methods of only 17 per cent of the firms. It may be of interest to note that the method for selection used in the largest percentage of firms was that of interviews, these being employed in 93 per cent of the firms.

Two years ago, Hubbard⁴ conducted a survey of employment methods used by various state governments. Questionnaires were sent to each of the 48 states and replies were received from 36. Of the 36 states replying, 7 had civil service commissions, 4 reported some form of centralized personnel agency other than a civil service commission, and 25 were without any centralized personnel agency. The results obtained are shown in Table IX. It will be noticed that intelligence tests are reported as being used in 7 of the 11 having some centralized personnel agency, but in only 1 in 25 of those without any centralized personnel agency. Obviously the use of mental tests in selecting personnel is related to the general development of progressive personnel methods.

The chief purpose of mental testing in vocational selection is to predict the fitness of prospective applicants

³ Scott, W. D., Clothier, R. E., and Mathewson, S. B., *Personnel Management*, A. W. Shaw Company, Chicago, 1931.

⁴ Hubbard, Henry Furness, *A Study of Objective Employment Tests in the New Jersey Civil Service*, unpublished thesis, 1934.

for the work for which they are applying. Fitness for work refers to that combination of traits and abilities which will enable the individual to maintain a satisfactory output of work with a minimum of expenditure of energy and with the least evidence of maladjustment to the job. Fitness for work is determined primarily by four factors: (1) *proficiency, knowledge, or skill*; (2) *capacity, competency, or general ability*; (3) *personality and character*; and (4) *interest in the work*. General mental tests or intelligence tests have a place in measuring the second of these factors, and it is the measurement of this factor with which we are concerned in the present discussion. The others of these factors will receive discussion elsewhere in the book.⁵

TABLE IX
EMPLOYMENT METHODS USED BY STATES

<i>Employment Method Used</i>	—NUMBER OF STATES—	
	<i>With central- ized personnel agency</i>	<i>Without centralized agency</i>
Test of knowledge of duties.....	9	1
Intelligence tests	7	1
Aptitude tests	6	1
Oral interviews	10	12
Experience qualifications	10	8
Educational qualifications	8	4
Medical examinations	8	2
Physical examinations	5	1
Recommendations	6	8
Application blanks	11	4
Number of states	11	25

Two types of studies have been of particular importance to the utilization of mental tests in personnel selection. The first is the study of intelligence levels of vari-

⁵ See Chapters XI, XIV, and XIX.

ous occupations. It seems quite obvious that if we are to utilize mental tests for selecting employees we should know what levels of intelligence are suitable for the various occupations. One of the most comprehensive studies of the intelligence level of men in various occupations is that made from the results of the Army Intelligence Tests. From the occupations of the men who were drafted in the Army, it was possible to make a distribution of the Army intelligence ratings and to compute the median scores for a great many occupations. Fryer later worked over the data afforded by the Army study, amplifying it in some instances by his own testing, and constructed a table showing the average scores and range of the middle 50 per cent of scores for a large number of occupations. Table X is quoted from Fryer's study. Such studies as these are very valuable, but they do not answer all the questions that we need to know about intelligence levels in occupations. Most of the studies show only the situation as it exists at the present time. The results are obtained by measuring the intelligence of persons on the job at the time of the study. They do not take into account the fact that there may be many individuals on the job who are too intelligent for the work, and many who are too low in ability for satisfactory work. At best, such studies can be only rough guides for setting up critical points below which or above which we would not desire to select employees.

The second type of study of immense importance to the utilization of mental tests in the selection of employees is the study of degree of relationship between mental ability and success on the job. Most of these studies have been done in terms of correlations between the intelligence test scores of employees who are selected to fill positions, and some measurement of their success on

TABLE X
INTELLIGENCE LIMITS FOR VARIOUS OCCUPATIONS*

<i>Army Alpha Score Average</i>	<i>Range of Middle 50% of Scores</i>	<i>Occupation</i>
161	110-183.....	Engineer (civil and mechanical)
152	124-185.....	Clergyman
137	103-155.....	Accountant
127	107-164.....	Physician
122	97-148.....	Teacher (public schools)
119	94-139.....	Chemist
114	84-139.....	Draftsman
111	99-163	Y.M.C.A. Secretary
110	80-128	Dentist
109	81-137.....	Executive (minor)
103	73-124	Stenographer and typist
101	77-127.....	Bookkeeper
99	78-126.....	Nurse
96	74-121.....	Clerk (office)
91	69-115.....	Clerk (railroad)
86	59-107	Photographer
85	57-110.....	Telegrapher and radio operator
83	64-106	Conductor (railroad)
82	57-108	Musician (band)
81	59-106 . . .	Artist (sign letterer)
81	60-106	Clerk (postal)
81	57-109.....	Electrician
80	62-114	Foreman (construction)
80	56-105	Clerk (stock)
78	54-102	Clerk (receiving and shipping)
78	61-106	Druggist
77	59-107	Foreman (factory)
75	56-105	Graphotype operator
74	53-91.....	Engineer (locomotive)
72	54-99	Farrier
70	46-95.....	Telephone operator
70	44-94	Stock checker
69	49-93.....	Carpenter (ship)
69	48-94.....	Handyman (general mechanic)
69	46-90.....	Policeman and detective
68	51-97.....	Auto assembler
68	47-89.....	Engineman (marine)
68	42-86	Riveter (hand)
67	50-92.....	Toolmaker
66	45-92.....	Auto engine mechanic
66	45-91	Laundryman

* From Fryer, D., "Occupational Intelligence Standards," by permission of *School and Society*.

TABLE X (Continued)

INTELLIGENCE LIMITS FOR VARIOUS OCCUPATIONS

<i>Army Alpha Score Average</i>	<i>Range of Middle 50% of Scores</i>	<i>Occupation</i>
66	49-86	Gunsmith
66	44-88	Plumber
66	44-88	Pipefitter
65	44-91	Lathe hand (production)
65	43-91	Auto mechanic (general)
65	43-91	Chauffeur
65	42-89	Tailor
65	44-88	Carpenter (bridge)
64	43-88	Lineman
63	40-89	Machinist (general)
63	46-88	Motorcyclist
63	41-86	Brakeman (railroad)
62	31-94	Actor (vaudeville)
61	40-85	Butcher
61	44-84	Fireman (locomotive)
61	39-82	Blacksmith (general)
60	38-94	Shop mechanic (railroad)
60	36-93	Printer
60	40-84	Carpenter (general)
59	40-87	Baker
59	39-83	Mine drill runner
59	38-81	Painter
58	37-85	Concrete worker
58	40-83	Farmer
58	37-83	Auto truck chauffeur
58	37-82	Bricklayer
57	41-81	Caterer
57	39-71	Horse trainer
56	38-76	Cobbler
55	35-81	Engineman (stationary)
55	34-78	Barber
55	35-77	Horse hostler
52	38-96	Sales-clerk
52	33-74	Horseshoer
51	31-79	Storekeeper (factory)
51	26-77	Aeroplane worker
51	31-74	Boilermaker
50	33-75	Rigger
50	30-72	Teamster
49	40-71	Miner (general)
48	21-89	Station agent (general)
40	19-67	Hospital attendant
40	19-60	Mason
35	18-62	Lumberman

TABLE X (Concluded)

INTELLIGENCE LIMITS FOR VARIOUS OCCUPATIONS

<i>Army Alpha Score Average</i>	<i>Range of Middle 50% of Scores</i>	<i>Occupation</i>
35	19-57.....	Shoemaker
32	16-59.....	Sailor
31	20-62.....	Structural steel worker
31	19-60.....	Canvas worker
30	16-41.....	Leather worker
27	19-63.....	Fireman (stationary)
27	17-57... ..	Cook
26	18-60.....	Textile worker
22	16-46.....	Sheet metal worker
21	13-47.....	Laborer (construction)
20	15-51.....	Fisherman

the job. A great many of these studies have been made on clerical and stenographic workers. Results reported by Bills⁷ seem typical of these studies. She found the correlation between intelligence and job success with 133 clerical workers to be +.22. She noted that those making the lowest scores tended to leave and that those making the highest scores tended to go into higher types of jobs. The same investigator studied intelligence in relation to selection of stenographers and typists.⁸ She expresses the opinion that the general intelligence test is the most efficient among a number of tests which she used. In her study, those individuals who were working as secretaries attained an average score on her test of 144; stenographers rated as "good," 110; stenographers rated as "getting along," 65; and stenographers failing in their work, 63.

⁷ Bills, M. A., "Relation of Mental Alertness Test Score to Positions and Permanence in Company," *Journal of Applied Psychology*, Vol. VII, pp. 154-156.

⁸ Bills, M. A., "Methods for the Selection of Comptometer Operators and Stenographers," *Journal of Applied Psychology*, Vol. V, pp. 275-283.

Hubbard reports typical studies of relationship between intelligence tests and success in certain jobs in the public service. He studied employees in the New Jersey Civil Service jurisdiction holding jobs as prison and reformatory officers, state patrolmen, firemen and bank examiners. He reports the correlations between success

TABLE XI

VALIDITY OF VARIOUS EMPLOYMENT METHODS

<i>Recruiting Method</i>	<i>Validity Coefficient</i>
Prison and Reformatory Officer:	
Final Average (Exclusive of Veterans' Preference)47 ± .06
Medical and Physical Test20 ± .07
Oral Interview24 ± .07
Experience and Training05 ± .07
Intelligence Test34 ± .07
Objective Duties Test51 ± .06
Patrolman:	
Final Average (Exclusive of Veterans' Preference)51 ± .07
Medical and Physical Test25 ± .09
Experience and Training09 ± .09
Intelligence Test50 ± .07
Objective Duties Test49 ± .07
Fireman:	
Final Average (Exclusive of Veterans' Preference)44 ± .08
Medical and Physical Test15 ± .09
Experience and Training17 ± .09
Intelligence Test31 ± .09
Objective Duties Test33 ± .08
Bank Examiner:	
Final Average (Exclusive of Veterans' Preference)58 ± .09
Oral Interview22 ± .13
Experience and Training36 ± .12
Free Answer Test26 ± .12
Intelligence Test (used for only one grade)28 ± .12
Objective Duties Test59 ± .09

on the job for these employees and various selecting methods as shown in Table XI. It will be noticed that in most of these instances the best single method of selecting employees is either by the intelligence test or by

the objective duties test. In one of these four types of positions, the mental test alone gives almost as high relationship with success on the job as is given by consideration of the average of all the factors studied.

Under the guidance of O'Rourke, the United States Civil Service Commission has developed a series of mental (general adaptability) tests, and O'Rourke^{*} has reported satisfactory use of these as part of the method of selecting applicants for a considerable number of federal government jobs.

TABLE XII

THE PROPORTION OF SUCCESSES AND FAILURES AT EACH MENTAL AGE LEVEL FOR 413 PACKING JOBS

<i>Mental Age</i>	<i>Total Number of Cases at Mental Age</i>	<i>Number Successes at Mental Age</i>	<i>Number Failures at Mental Age</i>
5 to 5-11	1	1	0
6 to 6-11	5	4	1
7 to 7-11	13	13	0
8 to 8-11	48	46	2
9 to 9-11	67	65	2
10 to 10-11	82	78	4
11 to 11-11	58	57	1
12 to 12-11	53	52	1
13 to 13-11	42	42	0
14 to 14-11 .. .	20	19	1
15 to 15-11	9	9	0
16 to 16-11	11	11	0
17 to 17-11	3	3	0
18 to 18-11	1	1	0

It is only fair to point out that despite such studies as we have just mentioned, intelligence is not found to be an important factor and mental tests are not found to be even moderately related to success in some jobs. This is particularly true of various low-grade factory jobs. A

^{*} See *Annual Reports* by L. J. O'Rourke, Director, Research Division, United States Civil Service Commission, 1923 to present.

typical study of this sort is one by Unger and Burr.¹⁰ Table XII shows the proportion of successes and failures among 413 persons doing packing jobs in an industry. Success is about as likely at the low mental ages as at the high ones.

It seems that we may summarize the values of mental tests in the selection of employees somewhat as follows: For most jobs at the professional, technical, or office level, intelligence, or general mental ability, seems to be a significant factor. It may be significant in the sense that a minimum level is required and that success above this is dependent on other factors; it may be significant in that degree of intelligence is directly proportional to degree of success all along the line; or it may be significant because of its relation to promotional possibilities in the organization. For most jobs of the routine factory type, there seems to be relatively little relationship between mental level and success. If there is a minimum level below which success is impossible, it is so low that the problem arises too seldom to justify mental measurements of all applicants. It should be emphasized that in personnel problems general mental tests should be supplemented by other measurements of human traits and capacities. Intelligence tests and intelligence test studies in industry have in many instances pointed the way to development of special vocational tests and aptitude tests.

IV. Uses of Mental Tests with Delinquents and Criminals

Long before the development of any quantitative methods of measuring human traits, those who dealt

¹⁰ Unger, E. W., and Burr, E. T., *Minimum Mental Age Level of Accomplishment*, University of the State of New York, Albany, 1931, p. 108.

with law-breakers were interested in the makeup of the criminal. Hence, with the advent of intelligence tests, criminals were among the early special groups studied. The studies have aimed at answering several questions about the criminal. Is he always, or commonly, mentally deficient? Is the mental defective always a potential criminal? Is there a relationship between type of crime and mental deficiency? What relationship exists between intelligence and possibilities of rehabilitation of the criminal? Not all the problems are yet fully answered, but the application of measures of intelligence has aided in the answering of some of them.

The statistical results. Most of the studies that have been made are based upon the testing of convicted delinquents or criminals in various types of institutions, such as reformatories, industrial schools, jails, prisons, penitentiaries. Most of the results have been expressed in terms of either the percentage of feeble-mindedness found among the criminal groups or the average intelligence of the criminal group. Most often the tests applied have been either the Binet Test or the Army Intelligence Tests. The studies have shown widely different results. Percentages of feeble-minded range from below 10 to over 90 per cent. Pintner lists 42 studies of feeble-mindedness among delinquent children and 13 of delinquent adults. The highest percentage of feeble-mindedness (93 per cent) was reported by Hill and Goddard in 1911. The lowest reported is 7 per cent, found by Miner in 1918 and Healy in 1922, both studying juvenile delinquents. Average percentages for all studies are about 30 for both children and adults. For the more recent studies, which we may consider somewhat more reliable, averages are closer to 20.

The extreme variations of the studies are probably to be accounted for by such factors as type of criminal tested, locality in which tests are made, type of mental test used, and representativeness of cases. The variations make it difficult to arrive at definite conclusions, but we can be fairly certain that feeble-mindedness is more prevalent among criminals in general than among the population at large. (Percentage among the population at large is near one per cent).

There have been several studies of the relation between type of crime and intelligence. Table XIII is quoted.

TABLE XIII

PERCENTAGE OF VARIOUS CRIMINALS
WHO ARE MENTALLY INFERIOR

<i>Crime</i>	<i>Per Cent Inferior (Below C, as Used in Army Tests)</i>
Fraud	22.0
Force	30.6
Statutory	31.0
Thievery	31.8
Physical injury	36.9
Dereliction	43.1
Sex	47.6

from a study by Murchison ¹¹ to show the types of differences usually found. Almost every type of crime seems to involve all degrees of intelligence. There is a tendency, however, for crimes against person and crimes of brutality to be committed by individuals of lower intelligence than crimes against property involve.

¹¹ Murchison, C., *Criminal Intelligence*, Clark University, Worcester, Mass., 1926.

V. The Use of Mental Tests in Various Theoretical and Research Problems

There are many psychological problems of theoretical importance or of only indirect practical bearing upon the problems of school, industry, or society that have been furthered or indeed made possible by the development of mental measurements. We mention a few of these.

1. **Studies of mental growth.** Two questions about mental growth have long interested psychologists. At what rate does mental growth take place at various points throughout the mental growth span—is the rate uniform or does it vary at different ages? At what age is mental maturity reached and growth stopped? Before the advent of mental tests, no accurate method was available for the study of these questions. Now a wealth of material has been collected in the form of mental test records for various ages. These records have generally shown increasing mentality up to an age somewhere near twenty, or somewhat above; furthermore, they have generally shown a decreasing rate of growth as age progresses. Thorndike has pictured the general nature of this growth as in Fig. 6.¹² Yet the questions about mental growth are far from settled at the present time. Pintner aptly points out the difficulties:

We must note, here, the great difficulty of arriving at any definite conclusion from such studies. If we take the average scores at each age as indicative of mental growth, we make two assumptions: first, that the score units are equal at all levels of the test, which is very doubtful; second, that the selection of children

¹² Thorndike, E. L., and others, *The Measurement of Intelligence*, Bureau of Publications, Teachers College, Columbia University, New York, 1927, p. 466.

at each age is the same, which is also very doubtful. The real curve of mental growth will not be obtained until we have a measure of intelligence with equal units at all levels and until we have repeated annual measures of a great number of cases.¹³

2. Studies of the constancy of relative mental ability.

Is a person's relative mental ability the same throughout life? Is the child of average I. Q. also the man of aver-

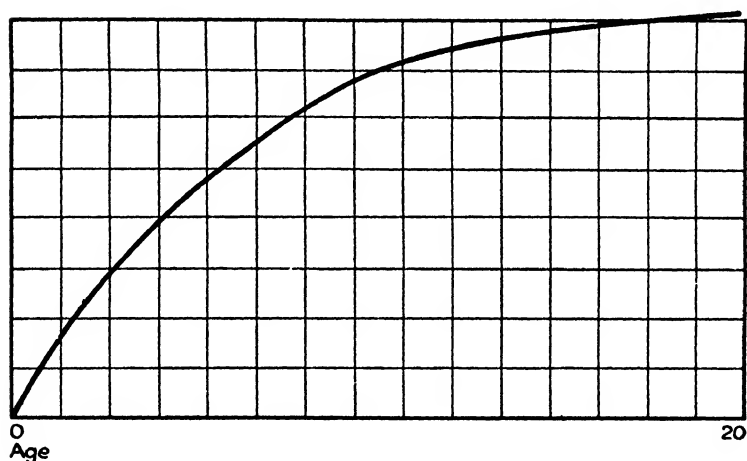


Fig. 6.—Mental Growth Curve.

age intelligence, or may relative ability change? These questions are of vital importance to our utilization of mental test results. If relative ability is likely to change, then our predictions on the basis of mental testing can be of only immediate value; remote or long-time predictions would be subject to great inaccuracies. Answers to the questions are contained in retests of individuals made considerable time after original tests. Generally

¹³ Pintner, R., *Intelligence Testing*, Henry Holt & Co., New York, 1931, p. 81.

such studies have shown surprisingly constant relative abilities of the testees. I. Q.'s (one of our commonest ways of expressing relative ability) remain closely the same from year to year. Typical studies concerning the constancy of I. Q.'s are shown in Table XIV, a summary taken from Freeman.¹⁴

TABLE XIV

MEASURES OF THE VARIATIONS IN THE I. Q. ON
RETESTING AS FOUND IN SEVERAL TYPICAL STUDIES

Author	No. Cases	Percentage Differing 10 Points or More	Limits of Middle 50 Per cent	Average Change	Coefficient of Correlation between Two Tests
Terman	435	.15	-3.3 to +5.7	4.5	.93
Rugg & Colloton.	137	.12	-2.3 to +5.6	4.7	.84
Garrison	468	.085	$\left\{ \begin{array}{l} -2 \text{ to } +4 \\ -3 \text{ to } +4 \\ -3 \text{ to } +5 \end{array} \right.$	5.4	.88
Rugg, L. S.	114	...	-1.2 to +1.9	3.1	.95

3. **Studies of heredity.** Since Galton's classic study of the inheritance of mental traits, described in his book *Hereditary Genius*, psychologists have pursued the question of the influence of heredity in determining mental ability. Most of the recent studies have made use of the correlation method, and of some type of mental test for indicating amount of intelligence. If the coefficient of correlation between mental test scores approaches zero, there is no resemblance in ability among the indi-

¹⁴ Freeman, Frank N., *Mental Tests*, Houghton Mifflin Co., New York, 1926, p. 345.

viduals compared. Randomly selected pairs of unrelated individuals show zero correlations for intelligence test scores, since there is no factor making for a resemblance among such individuals. Among pairs of related individuals, higher correlations are found, and the closer the degree of blood relationship between the pairs studied, generally the higher is the correlation coefficient. Such studies are not absolute proof of the influence of heredity.

But since environmental factors in most of the studies have been very similar, the assumption that the great resemblance in mentality of closely related individuals, as compared with less closely related or unrelated ones, is due largely to heredity does not seem to be unreasonable. Studies such as made by Merriam, Hildreth, Holzinger, and Thorndike show correlations of mental test records for identical twins of about .90, for fraternal twins of .70, and for siblings of about .50. Contrast these with correlations of zero for individuals paired at random.

4. Other studies using mental tests. Many other types of studies make use of mental tests. The scope of this book does not permit a discussion of them all. A few may be mentioned as suggestions for those who may wish to pursue them further. They include racial comparisons on mental tests, studies of relation between mentality and social status, sex comparisons in mental ability, studies of relationships between mental measurements and various physical and personality measurements, and many studies aimed at the improvement of mental tests themselves.

PART III

MEASUREMENT OF APTITUDES

CHAPTER VIII

The Measurement of Special Talents

THE instruments for measurement in the field of special talent are few, owing mostly to the nature of the things to be measured and to the infrequency of occurrence in the same individual of a sufficient amount of the talent or interest in the talent, on the one hand, and of psychological technique on the other, to make him an efficient investigator. Seashore, a pioneer in this field, working on the measurement of musical talent, stated in this way the problem which he faced:

The scientific study of the artistic mind is a somewhat baffling undertaking. There are no substantial precedents; the available scientific data are extremely meager; by nature the artist himself is but little interested in the process of his mental dissection; and, after all, the varieties of artistic minds are legion. But the time is ripe for a vigorous application of the technique of psychological inventory to practical affairs, and the discovery and fostering of human talents is indeed both practical and practicable. The stress of war forced our army to adopt psychological methods for the selection and rating of the human energies of men for assignment to service and for promotion. When the best results are demanded in any occupation, haphazard procedure must give way to procedure on the basis of ascertained facts. When Music shall come to her own she will come to the musically gifted: to that end musical talent must be revealed and encouraged.¹

¹ Seashore, C. E., *The Psychology of Musical Talent*, Silver, Burdett and Co., New York, 1919.

I. The Measurement of Musical Aptitude

As an example of measurement in the field of special talents, we shall examine Seashore's measurement of musical ability. The first mention of such standardized tests as this was made by Seashore in 1901. Years of research followed, during which the plans for these tests were experimentally developed in the psychological laboratory. As a final result, a battery of standardized musical tests was made available in 1919.

The problem of measurement of musical talent consisted, first, of an analysis of the elements making up musical ability; then, the discovery of objective ways of measuring them; and, finally, the standardization of the means of measurement. Seashore states that an inventory of the "musical mind" includes elements that have to do almost wholly with the reception of various attributes of sound: pitch, intensity, duration, and extensity—a problem of musical sensitivity; and elements that depend upon more intellectual qualities—a problem of musical appreciation, memory, and imagination. The first of these two types of elements is the more fundamental in the measurement of native capacity. The other type is dependent upon the first; is more subject to modification; is more difficult of analysis into its separate components; and is more mixed with other intellectual traits, no more related to musical performance than to other things the individual does. Five of the six elements measured in the Seashore tests are of the sensitivity group; one has to do with musical memory.

1. Description of the test. The six parts of the Seashore test measure the following: (1) sense of pitch; (2) sense of intensity; (3) sense of time; (4) sense of rhythm; (5) sense of consonance; (6) musical memory.

These six parts of the test are described by Seashore as follows:

The sense of pitch. The sense of pitch is involved, not only in the hearing of melody and harmony, but also in the hearing of tone character in many complex forms. Pitch is the raw material of music. The function of the higher capacities, such as memory, imagination, and feeling, or playing and singing, is limited by the degree of sensitiveness to pitch. This becomes significant when we find, for example, that, according to actual measurement, one person may be two hundred times as sensitive to pitch as another of equal age, social standing, and general intelligence.

The sense of intensity. Then, we have the sense of intensity, which represents the capacity for appreciation of differences in strength of sound. This is basic for the hearing of musical expression and the appreciation of touch, and for modulation in intensity or loudness and volume.

The sense of time. The third elemental capacity is the sense of time. This is basic for all perception of rhythm and for rhythmic action. A limitation in this capacity for hearing time sets a corresponding limitation upon feeling, thought, and action.

The sense of rhythm. The sense of rhythm rests upon the sense of time, the sense of intensity, and mental imagery, but it requires in addition a number of affective and motor qualifications; thus a person may have a keen sense of time and intensity and still not have a pronounced sense of rhythm.

The sense of consonance. The sense of consonance is the simplest form of musical hearing which underlies the combination of tones, either simultaneous or successive, as in melody or harmony. This rests primarily upon a sense of pitch, but involves higher elements so that a person may have a keen sense of pitch and yet not be effective in the sense of consonance.

Musical memory. The need of a musical memory is self-evident. It is not merely a matter of recalling se-

lections. Memory enters intricately into all stages of hearing, feeling, and rendering of music. The learning process is one special aspect of memory. Each individual has a certain personal equation for capacity in rate and excellence of learning, and each of us has some apt preference for one kind of material or another. For a given activity, such as singing, sight reading, piano exercises, this may be expressed in the form of what is technically called a learning curve.

2. How the six qualities are measured. The six parts of the Seashore test are recorded on phonograph records and the test is taken by having the subjects listen to the records and answer questions which would indicate their ability to discriminate differences in pitch, intensity, time, and so forth.

The sense of pitch, in terms of pitch discrimination, is measured by means of a series of tuning forks used with resonators. These forks are tuned in a differential series in which the standard is a fork with a vibration of 435. In the whole series there are ten forks which vary from the standard by different degrees of pitch. In the test the standard fork and one of the other forks are sounded in quick succession in front of a resonator. The problem of the listener is to tell whether the second tone is higher or lower than the first, and in taking the test he simply records "H" or "L" for each combination as he listens to it from the phonograph record.

The sense of intensity is measured in terms of the least perceptible difference in intensity or loudness from the standard sound. In producing the sounds for the musical test an audiometer is used. Two sounds are made for each trial, the differences in loudness between the two being of varying degrees. The subject is required to state for each set whether the second sound is weaker or stronger than the first.

The sense of time is measured as perceptions of dif-

ferences in duration of time intervals. The arrangement for recording this test on the phonograph record is one of a synchronous motor with a time-sense attachment. The attachment is such that clicks are produced by a projecting lever. These clicks, which occur at varying time intervals, are transmitted into a telephone receiver. From this point they are transferred to the phonograph record. In taking the test, the subject must indicate which time interval is the longest. For example, if there are three clicks, he indicates which interval is longer, the first or the second. If there are four clicks, he indicates the first, second or third.

The rhythm test is somewhat similar to the test for sense of time. The same instrument and the same method are employed as in the time-sense test except that the intervals are arranged in rhythmic divisions, long and short; for example, two quarter-notes and a half-note. Four such measures may be given as the standard, and in the fifth, the half-note may be made too long or too short, the test being to determine the accuracy in the perception of the long intervals.

The sense of consonance is the capacity for hearing differences in consonance and dissonance. It forms the basis of the ability to judge aesthetic effect in combinations of tones. The Seashore test for this quality is a test of appreciation of consonance and dissonance with reference to two-clangs, a two-clang being consonant when the two tones tend to blend and to produce a relatively smooth and pure clang; and, conversely, the two-clang being dissonant when the two tones do not blend and do not produce a smooth sound. The records recorded on the phonograph consist of sets of two-clang sounds, and for each set the subject must indicate which one of the two sounds is the more consonant.

The test for musical memory consists of a span of from

three to six musical notes which is played a second time with one of the notes changed. The subject taking the test must indicate which of the notes is changed in the second playing. The use of varying lengths of spans of notes gives an opportunity for distinguishing different degrees of ability to remember the notes played.

3. Studies of the Seashore test. (a) *Reliability.* There has been considerable disagreement as to the reliability of the Seashore test. Probably the most extensive study of reliability reported is that of Larson.² This study was made in Lincoln, Nebraska, and Iowa City, Iowa. Altogether between 1,100 and 1,200 individuals were tested. These were distributed among fifth-, sixth-, seventh-, and eighth-grade school groups and an adult group of somewhat over 400. All the groups tested were composed of unselected individuals so far as musical ability and musical training were concerned. The tests were given under well-controlled conditions; all possible precautions were taken to eliminate the influence of extraneous factors. A preliminary trial or fore-exercise was given to insure understanding of the directions. Tests which clearly indicated that the directions were misunderstood were discarded. The tests were given to the students by the group method, one test being given daily. The same test was given a second time at the same sitting after a short period of relaxation.

The reliability coefficients of each test for the same sitting were found by correlating scores on the first test with those on the second test. This was done separately for all six measures for fifth-, sixth-, seventh-, and eighth-grade and adult groups. The Pearson product-moment

² Larson, Ruth Crewdson, *Studies on Seashore's "Measures of Musical Talent,"* University of Iowa Studies, Vol. II, No. 6, 1927.

method of correlation was used in this study. The reliability coefficients are summarized in Table XV.

TABLE XV

RELIABILITY COEFFICIENTS FOR SEASHORE TESTS

<i>Test</i>	<i>Average Correlation for Adults</i>	<i>Average Correlation for 5th, 6th, 7th, & 8th Grades</i>
Pitch845	.85
Intensity765	.81
Time705	.71
Consonance725	.52
Tonal Memory ..	.935	.87
Rhythm ..	.700	.51

The final results showed that the pitch and memory tests are highest in reliability. Intensity ranks next, while consonance and rhythm show the lowest reliability in this battery of tests. The author of the study states that "the pitch and tonal memory tests are shown to be very satisfactory for use at all levels; the intensity and time tests are adequate for group measurements; and the consonance and rhythm measures should be reserved particularly for group averages and school surveys."

In discussing the reliability of the battery of Seashore tests, Larson attempts to account for the differences found in the results of different investigators. The following two paragraphs are quoted from the study.

The differences found in the results of various investigators might be attributed to numerous uncontrolled factors. Some differences might be expected, since results have been offered in connection with studies made primarily for purposes other than determining the reliability coefficients of the tests. Factors which might influence reliability are motivation, controlled room conditions, reliability of the phonograph motor, number of preliminary trials before the test is actually

given, adaptation to ages, the number of tests given at one sitting, dishonesty of students in recording, and interpretation. The writer observed that the size of the group tested made a difference; the larger the group the more difficult it was to keep controlled conditions.

Another factor that might affect reliability is the skill of the experimenter. Often in various aspects of research attempts are made to use the Seashore tests for a certain study without adequate preparation in the psychology of music or without an adequate technique acquired from a routine of music testing. Contrary to a common impression, experience is necessary to get the best results with the Seashore tests. From the "Foreword" of Stanton's *Prognosis of Musical Achievement*, Dr. Howard Hanson, director of the Eastman School of Music, in evaluating the eight year testing program at that school states, "As a practical musician I have been convinced of their (the Seashore tests') efficacy. I should wish, however, to add my belief that such testing is only of value when undertaken by thoroughly trained psychologists under conditions where control of experimentation is absolute. The undertaking of such a testing program by inexperienced and undertrained persons could only be a calamity."

(b) *Validity.* The study of the validity of the Seashore tests has presented a difficult problem because it necessitates having available, for comparison with the test scores, some reliable criterion of musical ability. As is obvious in a field where measurement has not been attempted before, it is practically impossible to obtain such a criterion. The test results have been checked with various indications of musical ability by different investigators. Stanton made a study of music students at the Eastman School of Music over a period of four years, in which an attempt was made to compare the musical test scores with the length of time the students stayed in school. This investigator reports that "of 242 cases,

47% of the cases which tested below average continued in their work for a year; over one half of the low talent discontinued. 84% of the talent which was rated as above the average remained one year." If we can assume that lack of ability to pursue musical training is a major cause of dropping out of the school, these figures indicate some validity of the test results.

Larson attempted to measure the validity of the test by comparing the test records for groups of six levels of musical ability; namely, first-class musicians, semi-professional musicians, advanced amateurs, beginning students, adult non-musicians, and eighth-grade non-musicians. In general this investigator found an increasing difference between the first-class musicians and each group of lower rank. The differences between the last three groups, the adult non-musicians, beginning music students, and eighth-grade non-musicians, are not large. This is as would be expected. Two of the groups are unselected, and the beginning music students rank only slightly above average in musical ability. Comparisons can best be made by knowing the differences between the first-class musicians, the semi-professional musicians, the amateurs, and the three unselected groups.

Larson also studied validity of the test by comparing test scores with school grades in music and with music teachers' estimates of students' ability. These studies were based upon students in the fifth, sixth, seventh, and eighth grades. The correlations which were obtained are summarized in Table XVI.

(c) *Norms.* Norms are given for the six parts of the Seashore test for three levels of ability; namely, for fifth grade, eighth grade, and adult. One might wonder why these three levels have been selected for the statement of norms. Seashore states that these levels are selected be-

cause of their relationship to vocational guidance in music, this being particularly true of the fifth- and eighth-grade levels. He thinks that the fifth grade represents a good stage for a survey of the musical ability of pupils in school, for at this stage they are able to take a responsible attitude and it is early enough to start them

TABLE XVI

VALIDITY COEFFICIENTS FOR SEASHORE TESTS

<i>Test</i>	<i>Correlation with Grades in Music</i>	<i>Correlation with Teachers' Estimates of Musical Ability</i>
Pitch32	.31
Intensity19	.11
Time21	.17
Consonance37	.34
Tonal Memory47	.46
Rhythm20	.23

in a musical education in case it has been neglected up to that time. All who have good musical talent can be measured reliably at that age if the test is carefully done. At the eighth grade, according to Seashore, the child faces another turning point. Here some children transfer to the high school and trade schools and enter upon a new adjustment of studies marked by the beginning of elective studies. Others leave school to work, and the avocation for life is frequently chosen in this pre-adolescent stage. Seashore thinks that for both of these classes of pupils, the claims of music, particularly as an avocation, should be presented in the most attractive form and with specific knowledge of the natural endowment of the pupil for music.

4. Studies of the relationship between the musical aptitude test and various other factors. The Seashore

test has been studied for its relationship to various other factors which may have some influence in determining musical ability.

(a) *Inheritance of musical traits.* Stanton used the Seashore test in a study of six of the foremost musical families in America. She examined 85 individuals in all. On the basis of the test results she concluded that there is a tendency for the inheritance of musical traits.

(b) *Racial influences.* Some half dozen studies have been made with the objective of comparing Seashore test records of whites and negroes. Most of these studies show no appreciable differences between the two groups in natural musical capacity. There is a study by Johnson which shows a superiority of the negro in the sense of rhythm; a study by Lenoir shows a superiority of colored children in sense of time and sense of rhythm.

(c) *Training.* A number of studies have been made on the effect of training upon the improvement of Seashore test scores. The results of these studies indicate in general that the capacities measured are relatively elemental in that they do not improve to any great extent with training.

(d) *Intelligence.* Most of the studies that have been made on the relationship between Seashore test scores and intelligence show low positive correlations. Studies of this nature have been made by Fracker and Howard (university students), Hollingworth (children above 135 I. Q.), and Larson (elementary school pupils). These studies seem to confirm the view of Seashore, that music tests are not in any significant way tests of intelligence.

II. Measurement in the Field of Art

More recently standard tests have been worked out in the field of art. Most of them measure art judgment

or artistic appreciation through the medium of responses to pictures. Studies have indicated that this type of artistic capacity is basic to artistic productions in the fields of painting, sculpture, and the like, and is fairly closely correlated with skill and technique in artistic performances; hence such tests as we are about to describe can be considered to have value in the educational and vocational guidance of talented individuals. Probably the most carefully devised of these tests are the Meier-Seashore Art Judgment Test and the McAdory Art Test.

1. The Meier-Seashore Art Judgment Test. The principles underlying the construction of this test have been stated by the authors as follows:

The underlying principle of the test is that *aesthetic judgment*, resting upon fine discrimination, feeling, and insight, is basic to success in art, whether it be sculpture, painting, etching, or some form of applied art. The possession of it in a high degree is what separates the master from the crowd, or the artist who produces *effective art* consistently from the one who seldom does. *Aesthetic judgment is defined as the capacity for perceiving quality in aesthetic situations relatively apart from formal training.* It may be further characterized as the capacity to sense artistic arrangement and effective use of line, color, and mass in a composition; to recognize good proportion in a vase or architectural design; or to feel the rhythm or movement of any type of art. Now, it needs only to be pointed out that, having this capacity, the artist will be accurately critical of his own effort and will sense, more frequently than one who has it in lesser degree, how to turn his efforts to advantage. He will have a definite feeling of when to add or omit details when their inclusion or omission may make or mar a partly completed composition. Research into the methods of great painters indicates that this capacity, functioning particularly in the matter of selecting the best sketch to develop from a number of prelimi-

nary studies, was one of the most significant traits of the old masters.

Talent in art is composed of a score or more of such related factors. In attempts to measure and evaluate these, some have been found to correlate slightly with ability to draw and apparent success in artistic pursuits; others have shown practically no relationship. The position taken here is that none of them is of such crucial importance or as indispensable to ultimate success as aesthetic sensitivity. The test measures the key-capacity, regarding it as the most trustworthy and significant *index* to talent and to probable success in an art career. Almost any student of art with a reasonable ability to draw may have some promise of success; yet it would be unfortunate for such a person, handicapped with a low degree of aesthetic judgment, to struggle along in competition with others better equipped. At least it would profit him to take fresh stock of his general capacities and abilities in other lines.³

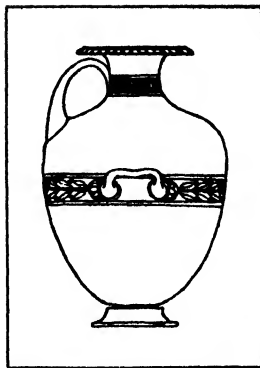
The test material consists of 125 pairs of pictures. The two pictures in each pair are nearly alike. They differ only in one respect, and the test blank indicates what that is in each case. The subject being tested records which picture in each pair is the better by encircling *L* for left picture or *R* for right picture. Pictures 9, 10, and 11 are reproduced in Fig. 7.⁴

Norms and distributions of scores on the test are published for junior and senior high school students, for art students, and for art teachers. The authors claim that validity of the test is indicated by (1) method of selection of the items; (2) differentiation in distributions of scores made by general groups as compared with groups

³ Meier, Norman C., and Seashore, Carl E., *The Meier-Seashore Art Judgment Test. Examiner's Manual*. Bureau of Educational Research and Service, State University of Iowa, Iowa City, 1930.

⁴ Reproduced by permission of Bureau of Educational Research and Service, State University of Iowa, Iowa City.

9



10



11



Fig. 7.—Samples from the Meier-Seashore Art Judgment Test.

L R 9 Location of the band.

L R 10 The arrangement of logs.

L R 11 Inclination of twig supporting bird.

of known artistic ability; (3) independence of scores of specific art training, shown by occurrence of high scores in general untrained groups; and (4) lack of significant correlation with general intelligence and other general ability measures. The selection of the items for the test was based upon several considerations. Many of the items are adaptations of "time-tested" pieces of art that have survived centuries of criticism. Final selection of the items was determined by, first, a favorable critical reaction on the part of 25 experts, and second, a 60- to 90-per-cent preference for the same item of the pair on the part of 1,081 subjects.

2. The McAdory Art Test. This test, worked out and standardized by Margaret McAdory Siceloff, consists of 72 plates, each presenting four illustrations which treat a single subject, or test item, in four different ways. The subjects of the plates include furniture and utensils, textiles and clothing, architecture, painting, and other graphic arts. Each plate calls for discrimination in one or more of the following art elements: shape and line-arrangement, massing of dark and light, color. Plates 5 and 23 are seen in Figs. 8a and 8b.⁵ Those taking the test are required to make a first choice, second choice, third choice, and fourth choice for each plate. The score is determined by the agreement of the testee's choices with those set up by consensus of a large number of competent judges selected from among artists, art critics, art teachers, etc.

According to the author's reports, the test seems to possess satisfactory reliability and validity. Scores on the test show increase with age up to about eighteen, differentiation between those with and those without artistic

⁵ Reproduced by permission of Bureau of Publications, Teachers College, Columbia University, New York.



A



B

Fig. 8a.—Samples from

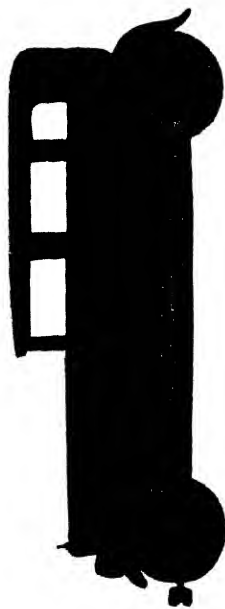


C

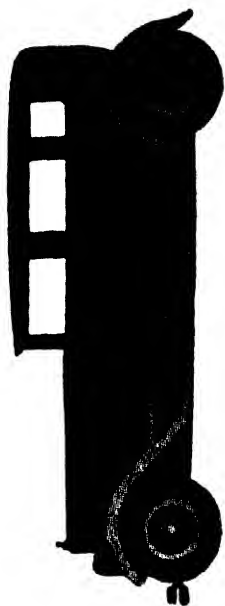


the McAdory Art Test.

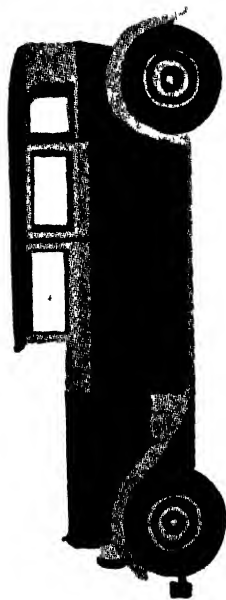
D



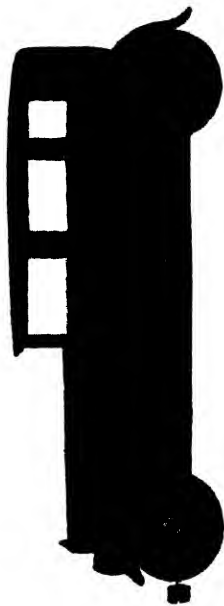
A



B



C



D

Fig. 8b.—Samples from the McAdory Art Test.

manifestations or training, and some advantage among females as compared with males of the same age and training. Some of these findings are indicated in Table XVII.⁶

TABLE XVII
NORMS ON THE McADORY ART TEST

<i>Based on Testing in New York City</i>	<i>Male</i>	<i>Female</i>
Zero ability, or score attainable by mere chance	68	68
Ability of an average 10-yr.-old child	103	114
Ability of an average 11-yr.-old child	110	123
Ability of an average 12-yr.-old child	116	132
Ability of an average 13-yr.-old child	122	139
Ability of an average 14-yr.-old child	127	145
Ability of an average 15-yr.-old child	132	150
Ability of an average 16-yr.-old child	136	154
Ability of an average 17-yr.-old child	140	156
Ability of an average 18-yr.-old child	143	158
Ability of an average adult	145	160
Ability of first-year student in art school	173	179
Ability of college graduate engaged in teaching	162	180
Ability exceeded by only 1% or fewer of adult population of the sex in question	202	220

⁶ Sicheloff, M., McAdory, and Woodyard, Ella, *Validity and Standardization of the McAdory Art Test*, Bureau of Publications, Teachers College, Columbia University, 1933, pp. 23 and 24.

CHAPTER IX

The Measurement of Mechanical Aptitude

WHAT are usually designated as mechanical aptitude tests grew out of the limitations of ability testing by the commonly used verbal intelligence tests. They were among the earliest of the special tests studied, Stenquist having begun work on the devising of his mechanical aptitude tests as early as 1915.

When work was started on measurement of mechanical aptitude, abstract intelligence was about the only ability which had been measured. Thorndike had already suggested a convenient division of general ability into abstract intelligence, mechanical intelligence, and social intelligence. And observation and experience without the aid of objective measuring devices had suggested that the second of these was sufficiently different from, and sufficiently uncorrelated in amount possessed with, abstract intelligence to make its measurement worth while. Stenquist was particularly interested in the possibilities which such measurements might open to the school children who measured relatively low on abstract tests but possessed considerable ability of a more concrete sort. This view is reflected in the following quotation from Stenquist: ¹

Of the relative importance of each of these two types of ability, readers must form their own conclusions. But it should be kept in mind that we are living in a

¹ Stenquist, J. L., *Measurements of Mechanical Ability*, Bureau of Publications, Teachers College, Columbia University, 1928, p. 89.

world that is dominated on every hand by every form of mechanical device and machine. Every moment of present-day life is influenced directly or indirectly by the products of mechanical skill and genius. Is it not important that ability in this field should be discovered and developed? Rather than merely to dismiss our apparently stupid pupils as low in what we now call general intelligence, and to relegate them to some convenient class, might not our time profitably be spent in disclosing other kinds of intelligence of which they may be possessed?

The question of "what knowledge is of most worth" will probably never be finally answered to the satisfaction of all. But it seems certain that as life becomes more and more complex, the world's tasks become more varied, and group inter-dependence increases, there is constant need for broader conceptions of what constitutes worth-while mental ability. We should recall that the history of the past century, as has often been said, could well be written in terms of the achievement of applied science and applied mechanical genius. Inventions of hitherto undreamed of significance, which have revolutionized or at least profoundly influenced the life of every nation on the globe, have sprung from this field of knowledge. And while the attempts to measure the mental abilities back of these forces, which are herein described, represent but crude beginnings, the importance of the task is stoutly maintained. Indeed, to explore, measure and adequately capitalize these capacities seems at least as important as doing the same for the more abstract type of intelligence required in academic school subjects. The discovery of special abilities has a two-fold significance and like the quality of mercy "is twice blessed": It not only opens the door of new promise to pupils, many of whom have been labelled as failures, but in doing so, it leads toward further contributions to society.

Tests of mechanical aptitude usually aim at measuring general aptitude in the management and manipulation

164 Measurement of Mechanical Aptitude

of things mechanical. They commonly imply a general knowledge of mechanical principles and usages, but do not presuppose any special trade skill. They draw upon inherited motor abilities and upon a certain amount of acquired familiarity with common things of a mechanical nature and upon mechanical interests. Applied to young people, high amounts of the traits measured by mechanical aptitude tests presumably represent potentialities of success in vocations and tasks requiring skill in the use of tools, or the understanding of and operation of machinery.

I. The Stenquist Mechanical Aptitude Tests

Stenquist developed two types of mechanical tests which have been widely used in schools and in industrial and business organizations, mainly in problems of vocational guidance and selection. He developed two Assembly Tests, known as Series I and Series II.² Each consists of ten common mechanical contrivances which have been taken apart and which must be assembled by the testee. The two series contain the following articles:

SERIES I

- | | |
|---------------------|-------------------|
| 1. Cupboard catch. | 6. Shut-off. |
| 2. Chain. | 7. Lock No. 1. |
| 3. Mouse trap. | 8. Push button. |
| 4. Hunt paper clip. | 9. Clothes pin. |
| 5. Bicycle bell. | 10. Wire stopper. |

SERIES II

- | | |
|-------------------------|------------------------------|
| 1. Sash fastener. | 6. Calipers. |
| 2. Rope coupling. | 7. Elbow catch. |
| 3. Defiance paper clip. | 8. Lock No. 2. |
| 4. Expansion nut. | 9. Expansion rubber stopper. |
| 5. Double-action hinge. | 10. Pistol. |

² Now made and sold by C. H. Stoelting Company, Chicago.

Score on the test depends upon speed and accuracy with which assemblies are made. Norms for various ages and grades have been prepared. A number of reliability coefficients, most of them above .60, are reported by the author.

Stenquist has made a number of studies of the validity of these assembly tests. He states that the best available criterion of general mechanical ability of the kind supposedly measured by the tests has been manual-training and science teachers' ranks of pupils. To safeguard as far as possible against untrustworthiness of ranks, Stenquist utilized for the most part only classes having two shop teachers, so that their rankings could be checked against each other. Correlations between the Assembly Test scores and ranks for various classes which he studied are given in Table XVIII. Most of these are high enough to indicate considerable validity for Stenquist's method; to indicate (so far as the criterion can be relied upon) that his tests really measure mechanical ability.

TABLE XVIII

CORRELATIONS BETWEEN SHOP-TEACHER RANK AND
SERIES I OF STENQUIST TESTS

7th and 8th grade boys in Lincoln school83
8th grade boys in New York City Public Schools80
8th grade boys in New York City Public Schools42
6th and 7th grade boys, Horace Mann School81
6th grade boys in Horace Mann School90
6th grade boys in Horace Mann School88

The relationship between the Assembly Tests and general intelligence has also been extensively studied by Stenquist. Published correlations are shown in Table XIX. Unless otherwise indicated, general intelligence is measured by the Army Alpha Test. The correlations

166 Measurement of Mechanical Aptitude

indicate very little relationship between general intelligence (abstract) and mechanical ability as measured by the Stenquist Tests.

TABLE XIX

CORRELATIONS BETWEEN STENQUIST ASSEMBLY TESTS AND GENERAL INTELLIGENCE

Camp Taylor, 109 unselected men323
Camp Devens, 107, foreign eliminated, but largely inferior cases35
Camp Lee, 76 unselected men30
Camp Lee, 30 men below 50 in Army Alpha .	.00 approx.
Camp Lee, 216 men low grade, individually ex- amined00 approx.
Camp Dix, 909 men, 303d Engineers, unselected	.51
Massachusetts School Feeble Minded, 30 cases with mental age32
Same group with officers' ratings25
For 100 7th- and 8th-grade boys, N. Y. Public Schools, between series I and composite intelli- gence score, made up of Haggerty, National, Otis, Kelley-Trabue and Meyers397
For same group, same tests, with Series II .	.338

Because the Assembly Tests are limited in usefulness on account of their cumbersomeness, expensiveness, and individual test nature, Stenquist devised another test of a pictorial nature. This test is known as the Stenquist Mechanical Aptitude Test, two tests, I and II, being available. All the problems of the tests are presented in the form of pictures. The pictures are of common mechanical devices, and the questions do not require special mechanical training or skill for their answering. A sample from Test I is shown in Fig. 9.³

Validity of the Picture Tests was studied by the same methods as were followed in the Assembly Tests, and by

³ From *Stenquist Mechanical Aptitude Test I*. Copyright by World Book Company, Yonkers-on-Hudson, New York. Reproduced by written permission of the publishers.

correlating the Picture Tests with the Assembly Tests. Correlations with shop ranks of students are similar to those found for the Assembly Tests. Correlations reported by Stenquist between the Picture Tests and the Assembly Tests average .67.

II. The Minnesota Mechanical Ability Tests ⁴

In 1930, Paterson and others published the results of an extensive investigation of mechanical aptitude tests at the University of Minnesota. The investigators

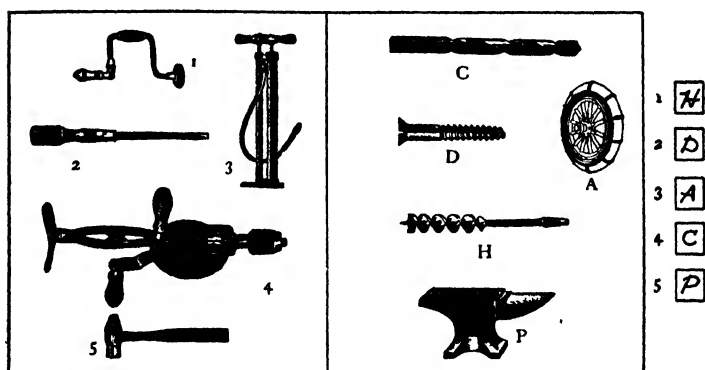


Fig. 9.—Sample from the Stenquist Mechanical Aptitude Test.
(Answered by matching letters with numbers.)

studied many types of tests which have been used to indicate mechanical ability, and arrived at batteries of tests which were demonstrated to have validity for measuring mechanical aptitude. Their tests were validated on boys in shop courses, using a combined (1) "quality criterion," based upon quality of work done in shop courses, (2)

⁴ Paterson, D. G., Elliot, R. M., et al.: *Minnesota Mechanical Ability Tests*, University of Minnesota Press, Minneapolis, 1930.

168 Measurement of Mechanical Aptitude

“quantity criterion,” based upon quantity of work done in relation to quality, and (3) “information” criterion, based upon objective information tests in the courses.

The battery of tests having the highest validity gave

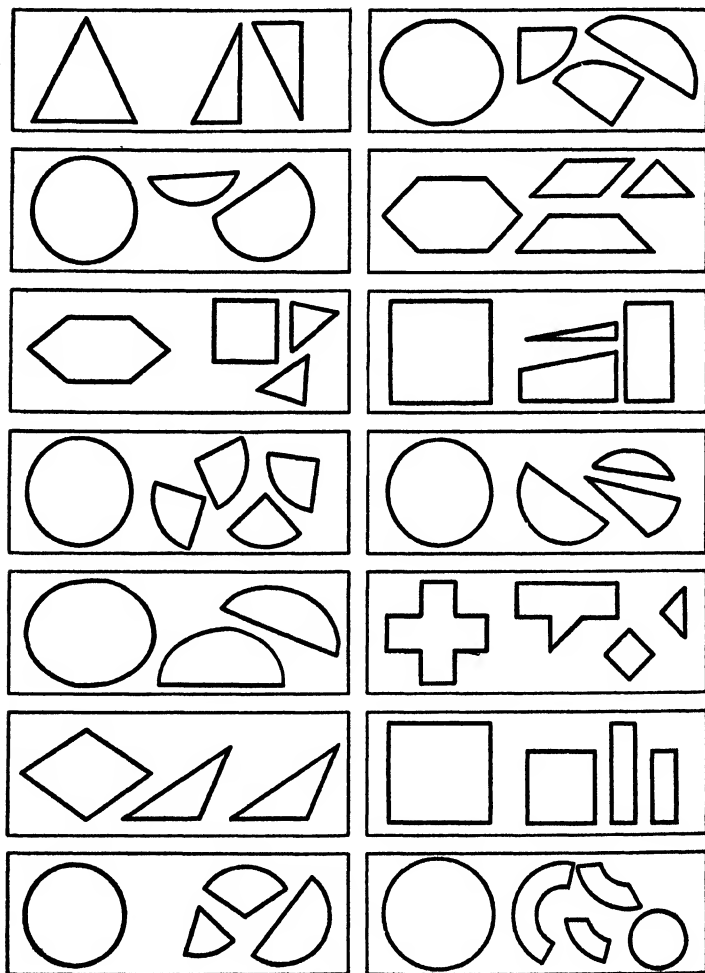


Fig. 10.—Minnesota Paper Form Board.

a correlation of .81 with combined criterion. This battery consisted of:

1. Minnesota Paper Form Board (see Fig. 10.)⁵
2. Academic Grades.
3. Minnesota Assembly Test (similar to Stenquist).
4. Boy's mechanical operations (based upon a questionnaire about home mechanical pursuits).
5. Interest analysis blank (to find out about mechanical interests).
6. Otis Intelligence Test.

The authors of the Minnesota Tests have studied the relationships between their tests and a considerable number of factors, as abstract intelligence, mechanical environmental influences, age, sex, etc. For discussion of these the reader is referred to the original publications.

III. The MacQuarrie Test for Mechanical Ability

This is a test of the pencil-and-paper sort. It consists of the following parts:

1. Tracing.
2. Tapping.
3. Dotting.
4. Copying designs, with a series of dots as guide.
5. Location—of dots in a group of letters.
6. Blocks—telling from a picture how many blocks a given block touches.
7. Pursuit—following a line through a tangle of lines and indicating where the line ends.

The MacQuarrie Test has the advantage of simplicity in administration and scoring. Its author reports a reliability of .90 and correlations with estimates of mechanical ability ranging from .48 to .80.

⁵ Reproduced by permission of the Marietta Apparatus Company, Marietta, Ohio.

Each picture below marked with a number is USED WITH a picture on the right marked with a letter. Look at picture No. 1, then look at the pictures on the right with letters and write the letter of the picture that is USED WITH it. Then find the picture that is USED WITH No. 2. The samples are done correctly. Picture C is USED WITH picture No. 1, so C is written after 1 on the line at the right. B is USED WITH 2. "Nail" marked A is USED WITH "hammer" marked 3, so write A after 3 on the line at the right.

NUMBERED

LETTERED

(WRITE ANSWERS HERE)



1	C
2	B
3	

[Below you will find questions relating to the pictures shown above.] In EACH square at the right of the questions, write a number or a letter to show which tools you would use to do what is asked. Where there are two squares be sure to write a number or a letter in each square. Pictures 3 and A are used to fasten a board to a box, so 3 and A are written in the squares at the right after question No. 1. Picture 2 is the correct answer for question No. 2. Pictures 1 and C are the correct answers for question No. 3, so you are to write 1 and C in the squares after question No. 3.

In each square on the right-hand side, write a number or a letter to show which tool in Figure 1 you would use:

(Write an answer in EACH square)

1. To fasten a board to a box
2. To tighten a nut
3. To fasten a door so as to use a padlock

Fig. 11.—Sample Questions from the O'Rourke Mechanical Aptitude Test. (Directions given in brackets have been slightly altered from the original.)

IV. The O'Rourke Mechanical Aptitude Test

O'Rourke's series of vocational guidance tests includes a mechanical aptitude test. This is mentioned particularly as an example of a mechanical aptitude test including more verbal material and somewhat more dependent upon language responses than such others as the Stenquist and MacQuarrie tests. O'Rourke's test contains two parts. Part One is a combination picture and verbal test, the nature of which is indicated in Fig. 11.⁶ Part Two is a verbal information test, in multiple-choice form, about mechanical things. Two sample questions are:

1. A carpenter drives a nail with a: (1) key (2) hammer (3) axe (4) screw (5) file Ans.-----
2. A carpenter drives a screw into wood with a: (1) hatchet (2) wrench (3) screw-driver (4) file (5) nut Ans.-----

Tests such as this are usually somewhat more difficult than the performance type, frequently show higher correlations with abstract intelligence (probably because of their verbal nature), and are less suitable for testing among uneducated groups or groups of low abstract intelligence. With many school groups, however, their usefulness is not below that of the less verbal test.

⁶ From O'Rourke *Mechanical Aptitude Test, Junior Grade*. Quoted by permission of the author.

CHAPTER X

The Measurement of Aptitude for Vocations and Professions

VOCATIONAL and professional aptitude tests are designed to predict ability to pursue successfully the various vocations, professions, or forms of training. They usually include measurement of the following qualities: (1) Intelligence in terms of the job; (2) Information either about the job or relative to the job; (3) Interest in the job; and (4) Any measurable special abilities requisite for the job. In this chapter we shall attempt to give an illustration of a test representing this field of measurement. We have selected the aptitude test for entrance to a professional school rather than the vocational aptitude test, since we shall have an opportunity to consider the latter in our discussion of psychological tests in industry.

The professional schools furnish excellent examples of the need for aptitude testing or of the need for testing fitness to profit by their instruction. From the standpoint of the student desiring to enter these schools, the training is expensive in both time and money. Much of the training is too specialized or technical to be of general value if the student is unable to complete the whole of the training. From the standpoint of the professional school, the cost of training is too high to eliminate the unfit by an attempt for a short time to train them. Misfits need to be eliminated or redirected before the training

is begun. An unusual opportunity for the use of aptitude tests is also created by the fact that the supply of "would-be professionals" greatly exceeds the capacity of the training schools and also is beginning to exceed the vocational need in the world. The problem as it presents itself, then, is one of selecting the most promising from the total number of applicants for admission to professional schools.

In the professional schools much less testing has been done than has been done in general intelligence testing of college students. This lag is probably due to several reasons. First of all, the task itself is inherently difficult, all the more so because the carrying out of a successful testing program in such a situation demands an individual trained both in the profession and in the technique of psychological testing. Another factor has probably been the lack of familiarity with tests and the general prejudice against them on the part of professional school administrative officials. However, there have been a few instances of significant work in this field. Ferson and Stoddard have published and standardized a Law Aptitude Test.¹ In the testing of entrants to engineering schools, considerable has been done at Stanford University by the use of a test which the author has termed a Scientific Aptitude Test.² Thurstone also has an engineering test of the aptitude type.³ Under the direction of the Association of American Medical Colleges, an extensive testing of medical school applicants has been carried out. Tests have also been given to teachers-in-

¹ Ferson, M. L., and Stoddard, G. D., "Law Aptitude," *American Law School Review*, 6:78, 1927.

² Zyve, D. L., *Stanford Scientific Aptitude Test*, Stanford University Press, 1929.

³ Thurstone, L. L., *Vocational Guidance Tests for Engineers*, World Book Co., Yonkers-on-Hudson, N. Y., 1922.

174 Vocational Aptitude Measurements

training.⁴ Since the testing program carried out in the selection of medical students has been the most extensive, we shall discuss it as a typical example of aptitude tests for use in professional schools.⁵

I. The Problem

Throughout the United States, there are 72 Class A medical schools which accept each year approximately 7,000 students for admission to their freshman classes. During the last few years there have been between 13,000 and 14,000 applicants per year. The problem, then, from the standpoint of the admission officials of the medical schools, is to secure the most promising 7,000 from the total number of applicants. The problem of selecting medical students is further emphasized by a study of the mortality of the graduating classes in the medical schools. In the past, more than one-fifth of all those who have begun work in the freshman year have been unable to complete the four-year course. Here is an ideal situation for development of an instrument of measurement which will assist in the selection of the best students for admission to the professional school. For a long time, of course, criteria have been set up for admission in these schools, and some of these criteria have been fairly exact in their selection.

II. The Development of a Medical Aptitude Test

The Scholastic Aptitude Test for Medical Schools, which we are to discuss in this chapter, was begun in

⁴ Moss, F. A., and others, *Teaching Aptitude Test*, Center for Psychological Service, Washington, D. C., 1927.

⁵ Material on Medical Aptitude Test is summarized from *Journal of Association of American Medical Colleges*, March 1930, Jan. 1931, May 1932, Jan. 1933, March 1934, and Jan. 1935. Quotations are made by permission of the Association.

1927. The test represented an attempt to devise an instrument which would indicate ability to pursue successfully a medical course, and which might be used as one of the determining factors in the selection of students for admission to medical school. During the school year 1928-29,* under the direction of the Association of American Medical Colleges, the first form of this test was administered for experimental purposes in 14 medical schools to about 1,000 students. The tests were given about the middle of the school year to students who were in the freshman medical-school classes. As these students progressed through the medical school, their records were followed closely in comparison with the aptitude test scores which they had made on entrance. During the following year, 1929-30, a second form of the test was similarly applied to an experimental group of about 5,000 freshman medical students, including those in most of the Class A medical schools in the country. Studies of these two experimental groups have been carried on up to the present time, some of the students having been followed through internships and State Board Medical Examinations. Some of the results of these studies will be indicated in our discussion. Following the experimental study of these aptitude tests for medical schools, the tests have been adopted as a requirement for admission to the medical schools and are now administered each year before applicants are considered for admission to freshman classes.

III. The Nature of the Aptitude Test

The test used in the studies being discussed has been developed in eight forms. These have contained the following parts: Scientific Vocabulary, Visual Memory, Memory for Descriptive Material, Pre-medical Informa-

176 Vocational Aptitude Measurements

tion, Learning and Retention of Material, Understanding of Difficult Printed Material, Ability to Follow Directions, Logical Reasoning, and Spelling of Scientific Words. All of the various parts of the test are given in terms of the scientific pre-medical terminology. The vocabulary test is selected from the terms met by the student in his pre-medical training in the basic sciences. The pre-medical information part of the test has also been drawn from the pre-medical courses in physics, chemistry, and biology. The learning parts of the test, including the visual-memory, memory-for-descriptive-material, and retention tests, are based upon a preliminary study of material selected from the field of medicine, unfamiliar to the student at the time that he studies it. His memory or learning ability in each instance is measured by his retention of the material studied.

The following samples from parts of the test will indicate more clearly its general nature:

SCHOLASTIC APTITUDE TEST FOR MEDICAL SCHOOLS (FORM 1)

TEST 1—SCIENTIFIC VOCABULARY

Directions. If a pair of words mean the same or nearly the same, encircle the *S*; if they mean the opposite or nearly the opposite, encircle the *O*:

- | | | | | | |
|---|---|---------------------------|---|---|-----------------------------|
| S | O | 1. anode-cathode | S | O | 6. relationship-correlation |
| S | O | 2. porous-impermeable | S | O | 7. deviation-aberration |
| S | O | 3. static-dynamic | S | O | 8. reduction-oxidation |
| S | O | 4. mechanistic-vitalistic | S | O | 9. objective-subjective |
| S | O | 5. vaporize-volatilize | S | O | 10. conduit-duct |

TEST 2—PRE-MEDICAL INFORMATION

Directions. If the statement is true, encircle the *T*; if it is false, encircle the *F*:

- T F 1. Limestone is an important source of calcium.
- T F 2. Ionization always accompanies electrolysis.
- T F 3. Hydrofluoric acid attacks glass.
- T F 4. Glycerol contains three hydroxyl groups.
- T F 5. The skull of vertebrates consists of a number of bones.
- T F 6. Hydrogen is generally prepared in the laboratory by the reaction of an acid and a metal.
- T F 7. Malaria is transmitted by a certain species of the *Anopheles* mosquito.
- T F 8. Standard conditions exist at a temperature of zero degrees Centigrade and a pressure of 760 mm. of mercury.
- T F 9. At the completion of a chemical reaction the properties of a catalyst remain unchanged.
- T F 10. Acetic acid is highly ionized.
- T F 11. The color of an object depends upon the wave lengths produced by it reaching the eye.
- T F 12. The salivary glands are endocrines.
- T F 13. The latent period in muscular contraction is the interval between the point of highest contraction and the beginning of relaxation.
- T F 14. The spinal cord is part of the central nervous system.
- T F 15. Thiocyanates contain sulphur.

TEST 3—VISUAL MEMORY

Directions. Below is the diagram which you studied at the beginning of the test. The various parts are numbered and the names appear at the right. Put in the blank before each name the number which designates that particular part.

178 Vocational Aptitude Measurements



----- Arch of Aorta
 ----- Inferior Vena Cava
 ----- Ascending Aorta
 ----- Pulmonary Artery
 ----- Thyroid Gland
 etc.

TEST 4—MEMORY FOR CONTENT

Directions. If the statement is true, encircle the *T*; if it is false, encircle the *F*:

- T F 1. The phrenic nerves supply the heart.
- T F 2. The coronary arteries supply the heart muscle.
- T F 3. In general, the venous apparatus is shown on the right side of the heart machinery.
- T F 4. The thyroid gland is situated between the thoracic aorta and inferior vena cava.
- T F 5. All blood leaves the heart from the ventricles.

TEST 5—COMPREHENSION AND RETENTION

Directions. Answer the following questions in accordance with the passage on "Speech Defects" which you read at the beginning of the examination. If the statement is true according to this passage, make a circle around the *T*; if it is false, make a circle around the *F*:

- T F 1. Psychoanalysis is used in treating speech defects due to emotional disturbances.
- T F 2. Some individuals are able to recognize objects but are unable to name them.

- T F 3. All paralyses of the vocal apparatus are caused by brain injuries.
- T F 4. Psychoanalysis is pointed out as being most effective in curing organic speech defects.
- T F 5. Injuries to the seventh nerve cause paralysis of certain of the muscles used in speech.

TEST 6—UNDERSTANDING OF PRINTED MATERIAL

Directions. Read the selection below and then answer the questions following it by writing the answers on the lines at the right. You may read the selection as many times as you desire and refer to it as often as necessary.

Planes and Diameters of the Pelvis

The planes of the pelvis are usually designated as (1) the superior strait, (2) the inferior strait, (3) the plane of the greatest, and (4) the plane of the least, pelvic dimensions.

The superior strait represents the upper boundary of the cavity, and is frequently spoken of as the pelvic inlet. It is somewhat oval in shape, with a depression on its posterior border corresponding to the promontory of the sacrum, and is sometimes described as blunt heart-shaped. It is bounded posteriorly by the promontory and alae of the sacrum; laterally by the linea terminalis; anteriorly by the horizontal rami of the pubic bones and the upper margin of the symphysis pubis. Strictly speaking, it is not a mathematical plane, since its lateral margins, as represented by the linea terminalis, are at a lower level than its central portion between the promontory and symphysis. Etc.

1. What is another name for the superior strait? -----
2. What diameter was named by Roederer? -----
3. Which of the four diameters of the superior strait furnishes the basis for estimating the size of the pelvis? -----

4. Which is the longest diameter of the superior strait? -----
5. What diameter is designated as the conjugata vera? -----

IV. The Validity of the Test

The validity of this aptitude test has been indicated principally by a study of the relationship between the test scores of students and their subsequent performance in the medical schools, their performance being based upon their medical school records. The students in the experimental groups who entered medical school in the fall of 1928 or the fall of 1929, and certain students tested later in pre-medical colleges, have been followed through their four years of medical school work. The average correlation between their four-year scholarship averages and their Medical Aptitude Test scores is stated as .59. This relationship is somewhat higher than that usually obtained between ordinary intelligence tests and scholastic records. The value of the test has also been studied from the standpoint of percentage of failures and percentage of high marks obtained in medical school by students of various test score levels. Such a study for one of the experimental groups is indicated in Fig. 12. The students were divided into ten equal groups on the basis of their scores on the aptitude test. Each bar of the chart shows the percentage of failures occurring in the four years and the average grade of the group. It is of interest to note that in the highest tenth no man has failed in any of the four years of the medical course, and the group shows an average grade for the four years of over 86 (an exceptionally high average for medical school grades). Eleven per cent of this group graduated with final medical school averages of 90 or higher. On the

other hand, at the end of the medical course, 60 out of 100 failed in the lowest tenth, and the final average for this whole tenth is only 75, just barely passing. No student shows an average as high as 90, and the majority of those who graduated show low grades and failures in some courses during the four years.

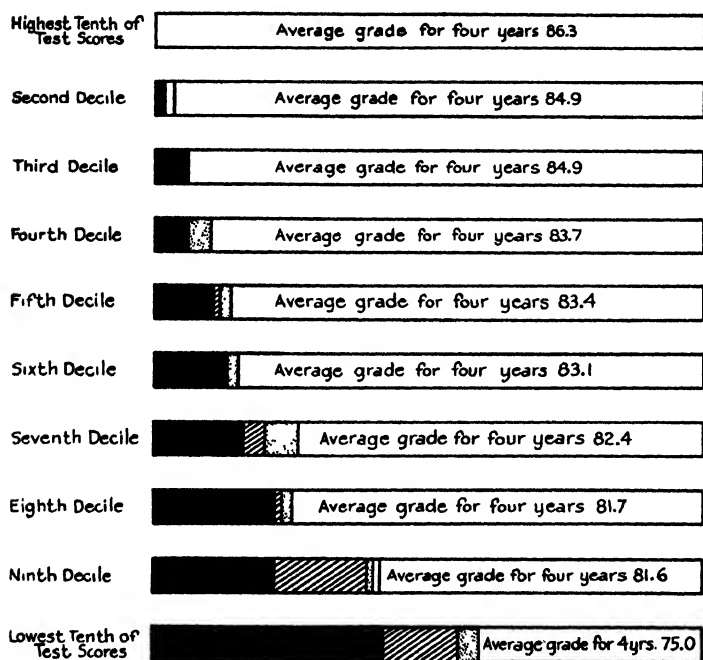


Fig. 12.—Distribution of Medical School Grades According to Medical Aptitude Test Scores.

Based on the distribution of medical school performances over the four-year course for these students, we may say that:

If a student has a score as high as the upper tenth tested:

182 Vocational Aptitude Measurements

- (1) The chances are 100 per cent that he will graduate from the medical school;
- (2) and the chances are 3 to 1 that he will average 85 or over for the whole four years.

On the other hand, if he is as low as the lowest tenth tested:

- (1) The chances are 60 out of 100 that he will not be able to graduate, because of failure to carry the work successfully;
- (2) and the chances are 9 to 1 that he will have an average below 85 if he does graduate;
- (3) or 2 to 1 that he will have an average below 80.

The validity of the Medical Aptitude Test has been further studied by comparing the success of certain of the students as internes in hospitals with their original Medical Aptitude Test scores. Such a study is reported for approximately 500 students interning in hospitals during the year 1932-33. Each student so tested was rated on his interne work by the superintendent or physician of the hospital in charge of internes. The ratings were made on a scale of one to five, "1" being a high rating and "5" a low rating. The interpretations of these ratings as given to the rater were as follows: "1" comes up to the best interne hospital has had; "2" is good, above average, but not equal to the best interne; "3" is equal to the average interne hospital has had; "4" is below the average interne, but better than the poorest hospital has had; "5" is among the poorest internes hospital has had. The relation between the aptitude test scores and the ratings as internes is shown in Fig. 13.

Another thing that might be of interest in our discussion of this aptitude test is the comparative study which

has been made of the test and other criteria often used in the admission of students to the medical schools. In any study of a new method of selecting students or employees, it would be logical to expect that a battery of criteria for admission or selection, the parts of which will supplement each other, would be the best. In the

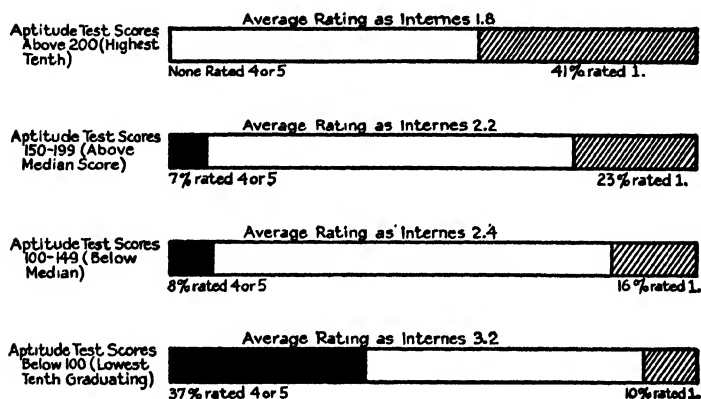


Fig. 13.—Distribution of Intern Ratings According to Medical Aptitude Test Scores.

medical schools there are various factors which have in the past been considered in admitting students. The most commonly used of these are: grades on pre-medical subjects; number of semester hours credit offered in pre-medical college work; personal interviews and character records; and in some instances, age at entrance. A comparative study of the value of each of these factors in predicting the failures occurring in medical schools has been reported for 1,000 medical students graduating in June 1932. The relative advantages of the five criteria indicated are shown in Fig. 14. The predictions of percentages of failures refer to percentages of actual failures which occurred in the group throughout their four years.

184 Vocational Aptitude Measurements

The percentage of good students excluded is indicated, since it is obvious that any criterion that is used to eliminate failures must be one which excludes as few good students as possible. These figures do not, of course, consider any possible good students eliminated by the

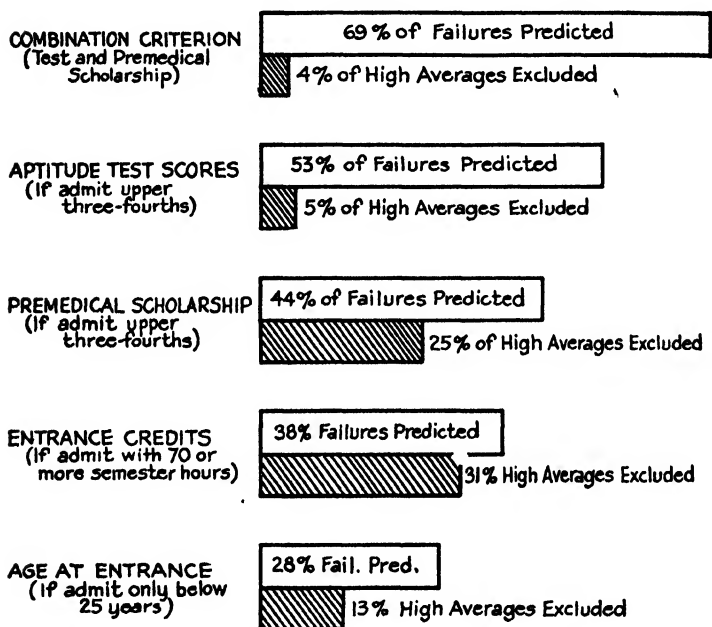


Fig. 14.—Comparison of Various Criteria for Admission to Medical Schools.

standards now in use (as “passing” pre-medical grades, or 60 semester hours of admission credits). Since students falling below these standards were not admitted, there was no chance to find out anything about them so far as medical school performance was concerned. Another factor studied and reported for a smaller group was the personal interview often given medical school appli-

cants. Thirty-three per cent of failures predicted has been indicated for this criterion of admission. Such studies as this indicate the relative validity of the test itself, and also emphasize the value of the test for use in conjunction with other criteria for selecting students, especially in conjunction with pre-medical grades.

Dr. Moss, who has acted as Secretary and Director of Study for the Committee on Aptitude Tests of the Association of American Medical Colleges since its foundation, in one of his reports for the Committee summarized the value of the aptitude test as follows:

We present the Aptitude Test as an additional criterion for selecting the best students for training in our medical schools. Studies of the validity of the various criteria which might be employed seem to indicate that the Aptitude Test is, in most instances, the best single criterion that we have; but we do not contend that in it we have a panacea. No criterion which we have been able to discover or set up is perfect; we can arrive at the best selection only by the wisest use of all the criteria available which demonstrate a real relationship to ability to pursue successfully a medical course.

Success in medical school undoubtedly depends on a complex set of factors, the most important of which may be listed as:

- (1) Innate ability or aptitude, for which our test is a very usable indication.

- (2) Previous preparation, a fair index of which is given by pre-medical grades.

- (3) Energy: It occasionally happens that a student has good ability and a satisfactory preparation but fails because of lack of energy. A physical examination may be of assistance in such cases.

- (4) Social adaptability or ability to fit in with fellow students and not antagonize instructors.

186 Vocational Aptitude Measurements

We have no reliable measure of this, but ratings by pre-medical instructors may be of service.

All of these four factors should be taken into consideration in selecting medical students. Efforts should be made to secure more reliable methods for measuring these factors.

CHAPTER .XI

The Measurement of Interests

ONE might question the inclusion of interests in this section of the book. No claim is made that interests are to be classed in the category of aptitudes or special talents. They are, however, usually somewhat specialized; at least, the infrequency of finding only general and broad interests makes the measurement of specialized interests desirable. Also, many of the same reasons for measurement of special abilities and aptitudes may be suggested as reasons for measurement of interests. We have, therefore, included the discussion of interests in this section dealing with measurement of aptitudes.

Interests have to do with our reactions to stimulus elements in our environment. The stimulus elements may be concrete objects, persons, ideas, techniques, or almost anything that we may think of as capable of calling out a reaction or a response in an individual. Fryer distinguishes between "subjective interests" and "objective interests"; and he also classes the various means that have been devised for measuring interests into these two groups. *Subjective interests* are dependent upon the feelings that accompany interest experiences. If the feeling is one of pleasantness, we term it an interest; if one of unpleasantness, an aversion. In the terminology of many of the measuring devices, interests are *likes* and aversions are *dislikes*. Between these are the indifferent

feelings of neither interest nor aversion. *Objective interests* are observable reactions to stimuli. If one walks down the street and stops to look intently at a sports exhibit in a window, we may say that he is showing a manifestation of interest that is objectively observable.

Vocational and occupational problems have probably carried the greatest weights in directing the attention of psychologists to the measurement of interests. Those who have given their attention to the problem have come to realize that the most helpful vocational guidance considers one's interest in following a vocation as well as his abilities for pursuing it; that ideal methods of selecting employees must take account of the fact that failures in work can occur from lack of interest as well as from lack of capacity or training. Fryer has emphasized another reason—a cultural reason—for the attention to measurement of interests:¹

The cultural importance of interest measurement . . . is only now becoming recognized. We have so long regarded abilities as the criterion of life's success that we have neglected the philosophy of happiness. To view interest measurement in true perspective we must start with an assumption different from the one current in this commercial age in which successful accomplishment is the criterion of measurement. Happy accomplishment is the foundation of a modern individualistic philosophy. Interest measurement is concerned with a distribution of interests which this philosophy assumes as the basis of happiness. The significance of interest measurement lies not in its relation to social efficiency, but rather in its measurement of a cultural development which is related to social happiness.

¹ Fryer, Douglas, *The Measurement of Interests*, Henry Holt & Co., New York, 1931, p. v.

I. The Methods of Measuring Interests

The chief methods that have been used in measuring interests are (1) inventories or questionnaires, (2) rating scales, (3) information tests, (4) free association tests. The first two of these fall in Fryer's classification of subjective measurement of interests. They represent an estimate by the individual of how he has reacted in the past or would react in the future to the stimuli of the test. The third and fourth methods are objective methods of measurement. These tests measure reactions to stimuli under standard stimulation conditions, without consideration of the estimate by the person being tested or by others.

1. **The inventory or questionnaire.** The interest inventory presents to the individual a list of items calculated to stimulate people to feelings of like or dislike, and asks him to estimate his feeling toward the items just as if they were stimulating him. The items may be occupations, books, magazines, people, amusements, or courses of study. Most frequently the estimated interests toward the items are to be recorded in terms of *Like*, *Dislike*, or *Indifference*. Finally, the interest inventory usually provides a method of scoring or summing the responses of the testee to indicate degree of interest in various broad fields, as an occupation, or a field of interest such as mechanical interest.

The best example today of a standardized interest inventory for general use with adults is the *Vocational Interest Blank* by E. K. Strong, first published in 1928. Very recently Strong has published a form especially adapted for women. The "Vocational Interest Blank" represents a revision and extension of an earlier inventory worked out by Cowdery. Together at Stanford Univer-

sity, Cowdery and Strong have conducted the most extensive studies available on measurement of interests. Strong's interest blank contains 420 items classified into eight parts, as follows: (1) Occupations, (2) Amusements, (3) School Subjects, (4) Activities, (5) Peculiarities of People, (6) Order of Preference of Activities, (7) Comparison of Interest between Two Items, (8) Rating of Present Abilities and Characteristics. Sample items from the blank are quoted to show its nature. The items are in most instances to be marked as Like (L), Indifferent toward (I), or Dislike (D).²

PART I. OCCUPATIONS

Actor (not movie)	L	I	D
Advertiser	L	I	D
Architect	L	I	D
Army Officer	L	I	D
Artist	L	I	D
Astronomer	L	I	D
Athletic Director	L	I	D
Auctioneer	L	I	D
Author of Novel	L	I	D
Author of Technical Book	L	I	D
etc.			

PART II. AMUSEMENTS

Golf	L	I	D
Fishing	L	I	D
Hunting	L	I	D
Tennis	L	I	D
Driving an Automobile	L	I	D
Taking Long Walks	L	I	D
Boxing	L	I	D
Checkers	L	I	D

² Strong, E. K., Jr., *Vocational Interest Blank*, Stanford University Press, Stanford University, California, 1928.

Chess	L	I	D
Poker	L	I	D
etc.			

PART III. SCHOOL SUBJECTS

Algebra	L	I	D
Agriculture	L	I	D
Arithmetic	L	I	D
Art	L	I	D
Bible Study	L	I	D
Bookkeeping	L	I	D
Botany	L	I	D
Calculus	L	I	D
Chemistry	L	I	D
Civics	L	I	D
etc.			

PART IV. ACTIVITIES

Repairing a clock	L	I	D
Making a radio set	L	I	D
Adjusting a carburetor	L	I	D
Repairing electrical wiring	L	I	D
Cabinet making	L	I	D
Operating machinery	L	I	D
Handling horses	L	I	D
Giving "first-aid" assistance	L	I	D
Raising flowers and vegetables	L	I	D
Decorating a room with flowers	L	I	D
etc.			

PART V. PECULIARITIES OF PEOPLE

Progressive people	L	I	D
Conservative people	L	I	D
Energetic people	L	I	D
Absent-minded people	L	I	D
People who borrow things	L	I	D
Quick-tempered people	L	I	D
Optimists	L	I	D
Pessimists	L	I	D

People who are natural leaders L I D
 People who assume leadership L I D
 etc.

PART VI. ORDER OF PREFERENCE OF ACTIVITIES

(Three lists of ten activities each. In each list the three enjoyed most and the three enjoyed least must be checked.)

PART VII. COMPARISON OF INTEREST BETWEEN TWO ITEMS

Street-car motor-				Street-car conduc-
man	()	()	()	tor
				Fireman (fights
Policeman	()	()	()	fire)
Chauffeur	()	()	()	Chef
Head waiter	()	()	()	Lighthouse tender
House-to-house				
canvassing	()	()	()	Retail selling
House-to-house				
canvassing	()	()	()	Gardening
Repair auto	()	()	()	Drive auto
Develop plans	()	()	()	Execute plans
				Delegate job to
Do a job yourself . .	()	()	()	another
Persuade others . . .	()	()	()	Order others

PART VIII. RATING OF PRESENT ABILITIES AND CHARACTERISTICS

	Yes	?	No
Usually start activities of my group . . .	()	()	()
Usually drive myself steadily (do not			
work by fits and starts)	()	()	()
Win friends easily	()	()	()
Usually get other people to do what I			
want done	()	()	()
Usually liven up the group on a dull day	()	()	()

Am quite sure of myself	()	()	()
Accept just criticism without getting sore	()	()	()
Have mechanical ingenuity (inventive- ness)	()	()	()
Have more than my share of novel ideas	()	()	()
Can carry out plans assigned by other people	()	()	()

(a) *The selection of items for interest inventories.* Evolution of the interest inventory shows two criteria that have been set up in the selection of items. One is a *sampling* criterion. No inventory could contain all the items of interest, or aversion, with respect to a particular interest group, so the aim has been to obtain an adequate sampling of the total possible items. A second criterion has been a *discrimination* criterion. The aim has been to select items that discriminate in the interest estimates assigned them between groups of people, occupational groups in particular, and to omit those items that are common interests or aversions of the different groups to be distinguished by the inventory. Such discriminative value of the items of an inventory can be discovered only through detailed tabulation of responses made by various groups of persons.

(b) *The scoring of interest inventories.* Strong's scoring of his *Vocational Interest Blank* exemplifies the most refined technique in the scoring of interest inventories. Scores may be obtained for interest in 30 different occupations. For each occupation, each possible answer receives a scoring weight dependent upon the difference in per cent of answers of the occupational group for which the scoring is being worked out, and a group of "men in

general" representing many occupations. Strong employs a special formula for working out these weights.³

The general validity of the weights assigned to items is reflected in the differentiation of the occupational group in question from "men in general" when scored for interest in the occupation. This may be illustrated by Strong's study of the occupation of personnel manager. When the interest inventories are scored with the per-

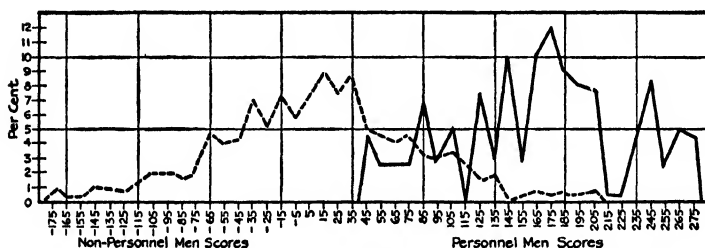


Fig. 15.—Distribution of Scores for Interest in Personnel Management.

sonnel managers' scoring key, the scores are distributed as shown in Fig. 15. The range of scores for the personnel managers is from 45 to 285, for the non-personnel group from -175 to 225. The overlapping by the non-personnel group is about 28 per cent. Seventy-two per cent of the personnel managers are distinguished from the non-personnel men, or, to state the situation another way: the chances are 72 in 100 of selecting personnel managers from non-personnel men by their interests.

Strong reports similar studies for his various other occupational groups. For example, chances in 100 of distinguishing the occupational groups from a general group of mixed occupations are: 95 for artists, 84 for ministers,

³ Strong, E. K., Jr., "Procedure for Scoring an Interest Test," *Psychological Clinic*, Vol. XIX, 1930, p. 63.

75 for lawyers, and 64 for engineers. Not all occupations that Strong has studied, however, can be differentiated to a degree that would make of practical value their interest measurement by his inventory. His "executive" group could not be distinguished to any practical degree from non-executive groups. The chances of selecting "executives" by interest score from a non-executive group was only 30 in 100.

2. **The rating scale.** In the measurement of interests, the rating scale has been relatively little used as compared with the inventory or questionnaire. In many instances, however, the rating technique would seem to be as applicable to the field of measurement of interests as to that of measurement of abilities. We may illustrate the technique in the measurement of interests by a procedure used by Cox in studying the interests of a group of genius children.⁴ She used a seven-point scale ranging from plus 3 to minus 3. Plus 3 was the rating assigned for interest of the highest degree; plus 2, interest considerably above average; and so on, to minus 3, which designated interest of the lowest degree. Her subjects were rated on seven aspects or types of interest, including intellectual interests, social interests, activity interests, breadth of distinct interests, breadth of related interests, intensity of a single interest, and intensity of two or more interests.

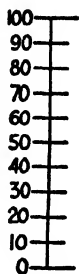
Kitson⁵ has also utilized the rating scale technique in some of his studies of vocational interests. He has asked his subjects to rate interest in their vocation on a numerical scale by comparing it with interest in other occupa-

⁴ Cox, C. M., "The Early Mental Traits of Three Hundred Geniuses," *Genetic Studies of Genius*, Vol. II, Stanford University Press, 1926.

⁵ Kitson, H. D., "Measuring the Interest of Teachers in Their Work," *Teachers College Record*, Vol. XXX, No. 28, 1928.

tions. His "vocation-to-vocation" rating scale of interests is reproduced below.

VOCATION-TO-VOCATION RATING SCALE OF
INTERESTS (KITSON)



Indicate, by making a check on this scale, the degree of interest you have in your occupation (not the present job, but the occupation itself). As the 100-degree point, think of that activity in which you would spend the major portion of your time if you had a million dollars and were not obliged to work. Then check the point on the right of the scale which denotes the degree of your interest in your present occupation.

3. Information tests. This method of measuring interests is based upon the assumption that knowledge and interest go together, that one will tend to inform himself about those things which interest him. Burt, who worked out an Interest Test of Agricultural Engineering on this basis, states the theory back of such measurement as follows:⁶

There is some ground for the assumption that if a person is interested in a certain field he will pick up information about it—will be more familiar with the terminology and with less obvious details that would presumably be overlooked by a person who lacked that interest. Consequently, an information test may give some indication of interest if the items are carefully selected.

We mention some examples of the use of this method of measuring interests.

⁶ Burt, H. E., *Employment Psychology*, Houghton Mifflin Co., Boston, 1926, p. 302.

(a) *Mechanical interest tests.* In our discussion of mechanical aptitude tests we noted that certain measurements in this field made use of information tests. Part II of the O'Rourke Mechanical Aptitude Test⁷ is essentially an information test, and is based upon the assumption that those with mechanical aptitude will be interested in mechanical things and will, therefore, acquire information above the average about such things. Tests in the field of mechanical interests utilizing the information test method were also developed for use in the Army during the World War period, largely through the efforts of Toops and O'Rourke. Their "General Trade Interest Test" consisted of a number of one-word answer questions calling for information in the general field of mechanical trades. These tests were found very useful in the placement of soldiers and in the selection of those desiring to enter the Army trade schools.

(b) *Social interest tests.* There are several instances in which information tests have been used as interest tests in the field of social measurements. The early form of the "Social Intelligence Test" (Moss, Hunt, and Omwake) made use of a Social Information Test on the assumption that breadth of interests is related to one's social intelligence and that this breadth can be measured by an information test. The material of this test was expressed in true-false form. The following are sample items:

- T F 1. The nickname of the Chicago Nationals is Red Sox.
- T F 2. A white tie should be worn with a tuxedo suit.
- T F 3. The term "right bower" is used in playing bridge.

⁷ See p. 171 for sample questions.

- T F 4. Coney Island is near New York.
T F 5. In hotels run on the European plan the charges include both room and meals.
T F 6. The composer of the opera "Faust" was Gounod.
T F 7. Membership in the American Automobile Association is limited to members elected each year.
T F 8. The Speaker of the House of Representatives is elected by members of the House from their own number.
T F 9. The Statler system refers to a chain of department stores.
T F 10. In federal elections for President the majority of votes cast in Tennessee are usually for the Democratic candidate.
T F 11. The Army and Navy football game is usually played at either West Point or Annapolis.
T F 12. Jane Cowl is a vaudeville actress.

Ream in 1921 devised a "Social Relations Test" based upon the same general assumptions that form the basis of the Social Information Test. His material is in multiple-choice form, such as:

2. In what organization is eleven o'clock of special significance? Elks—Odd Fellows—Masons—Knights of Columbus.
4. What is a caucus? A national political convention—An official county election—A meeting of politicians within a party—A secret political meeting in violation of the law.
5. What kind of race is a derby? Trotting—Pacing—Running—Hurdling.

(c) *Vocational interests tests.* An example of a test in this field is McHale's *Vocational Interest Test for College Women*. The test contains 247 questions in multiple-choice form divided among four types of vocational

interests: law, business, medical sciences, and homemaking.

(d) *Terman's information test of play interests.*⁸ This test was used by Terman in his investigation of play interests of gifted children as a part of his extensive study of superior children. It is a multiple-choice test covering various types of games—social, solitary, active, quiet, etc. Questions are such as:

1. You pick up jackstraws with a Magnet—Hook—Fingers.
2. A game where you look for something hidden is I-spy—Old Witch—Roly-Poly.
3. "Hearts" is played with Cards—Dice—Dominoes.

Terman gives norms based upon several hundred cases for normal and genius children of both sexes.

It is rather difficult to make any very conclusive generalization as to the value of information tests as a measure of interests. The information test possesses high reliability, in most instances considerably higher than for other methods of measuring interests. It has the advantage of being applicable to various types of interest testing. It can be adapted to testing general range of interests or to testing interests in specialized fields. Information tests as measures of interests are likely to fall short of the purposes for which they are devised in failing to sample extensively enough the field of testing, in measuring ability or experience rather than interests, and in putting too high a premium on abstract intelligence and verbal ability. The most carefully constructed tests of this nature, however, seem to possess a value

⁸ Terman, L. M., and others, "Mental and Physical Traits of a Thousand Gifted Children," Vol. I, *Genetic Studies of Genius*, Stanford University Press, 1925.

worthy of retaining the method in further studies of measurement of interest.

4. Association tests. Association tests of the type already described in Chapter VI have been used as a means of indicating a person's interests. These tests are based upon the assumption that a person with a particular kind of interest will be likely to respond to a stimulus word, in a large percentage of cases, with a particular kind of response indicative of his interest. Provided a long enough list of stimulus words is presented, such a test may give a fairly reliable indication of the general field of interest. The most extensive investigation of interests by this method of testing has been carried on by Wyman. She devised an interest test suitable for young people between the ages of eight and fifteen, for studying interest reactions in three fields: intellectual, social, and activity interests. By intellectual interest, she had in mind interest in "knowing—in getting the meaning of things." By social interest, she meant interest in persons. Activity interest she considered to be interest in doing things—interest in being the leader. Her association test consists of 120 stimulus words to be presented verbally with responses to be given in writing. The words have been balanced equally for responses in intellectual, social, and activity fields of interest. Her test list is reproduced on page 201. Either list of 60 words may be used for a short test.

The real value and the validity of an association test for interests depend upon the care and accuracy with which the scoring key is prepared. Wyman's method makes use of a technique similar to that used by Strong in assigning score weights to the various responses in his inventory. The various responses to the stimulus words are assigned score weights for each of the interest groups

by comparing the responses of a trial group *with* the interest with those of a group *without* it. For example, a particular response to a stimulus word receives a score weight for intellectual interest in proportion to the degree to which that response differentiates individuals of known intellectual interests from individuals known to be lacking in such interests. As is true in scoring the responses of the interest inventories, the working out of scoring

WYMAN STIMULUS WORDS FOR TESTING INTELLECTUAL,
SOCIAL, AND ACTIVITY INTERESTS⁹

1. summer	31. evening	1. night	31. sundown
2. easy	32. hard	2. simple	32. difficult
3. diamond	33. ring	3. gem	33. dress
4. tire	34. play	4. join	34. enjoy
5. dog	35. learn	5. control	35. need
6. fair	36. band	6. white	36. music
7. school	37. dark	7. college	37. black
8. help	38. platform	8. protect	38. stage
9. nature	39. pity	9. sky	39. watch
10. active	40. thrill	10. restless	40. excite
11. dream	41. idle	11. wonder	41. useful
12. shock	42. hero	12. fault	42. castle
13. joy	43. vacation	13. pleasure	43. holidays
14. dislike	44. master	14. detest	44. captain
15. nut	45. bat	15. paper	45. rod
16. go	46. fun	16. travel	46. mischief
17. angel	47. power	17. princess	47. rain
18. nice	48. interested	18. alone	48. interesting
19. water	49. fond	19. current	49. good
20. boy	50. trip	20. girl	50. journey
21. wish	51. make	21. desire	51. form
22. museum	52. yard	22. history	52. island
23. delight(ed)	53. aim	23. contented	53. try
24. work	54. fairy	24. train	54. giant
25. cave	55. exercise	25. adventure	55. game
26. pleasant	56. companion	26. happy	56. friend
27. house	57. career	27. marble	57. science
28. imagine	58. fire	28. invent	58. camp
29. range	59. like	29. country	59. prefer
30. admire	60. great	30. attract	60. grand

⁹ Wyman, J. B., "The Measurement of Interest," *Vocational Guidance Magazine*, 1929, Vol. VIII, p. 54.

methods for association tests of interests is a tremendous job.

Wyman's test possesses a satisfactory reliability, as shown by reliability coefficients between .80 and .90 when the test is scored for the three types of interests. The validity of the method has been studied by correlating the interest scores with teachers' estimates of dominant interests in pupils. Such correlations reported by the author of the test are generally between .45 and .65. These seem as satisfactory as most validity coefficients based upon personal estimates. At best, they furnish a questionable basis for validation, but it is difficult to obtain a more satisfactory criterion with which to compare the test scores.

Wyman has studied several questions bearing a relationship to the use of this testing device. She has studied the relationship between the interest scores and intelligence and achievement test scores. The former give an average correlation of .48 with intelligence test scores and an average correlation of .50 with Stanford Achievement Test scores. Working out partial correlations between interest scores and achievement scores holding intelligence constant, Wyman found correlations as follows: for intellectual interests, .49; for social interests, .18; and for activity interests, .03. The correlation of .49 for intellectual interests is three points higher than the correlation between intellectual interests and intelligence scores, which might show the importance of intellectual interests as well as ability in achievement. Wyman also made a study of the permanence of interests as measured by her test. The average correlation for the three fields of interests tested the second time, after a period of five years, is .28. It would seem that interests in these three fields lack any high degree of permanence

in school children. This is in substantial agreement with studies of permanence of interests using other methods of measurement.

The following quotation from Fryer¹⁰ seems a fitting conclusion to our whole discussion of the measurement of interests:

The present state of development in the measurement of interests corresponds to an early stage in the measurement of abilities. Twenty years ago there was one impressive scale for the measurement of abilities, the Binet-Simon Scale, and many other experimental tests. In the measurement of interests, today, there is one outstanding measuring scale, Strong's "Vocational Interest Blank," with its scoring keys for various occupations, and many other experimental devices which may prove satisfactory in the future.

There is, however, an important difference to be noted in this comparison of development in the two fields of measurement. The measurement of abilities had achieved an objective basis by 1908. Today the objective measurement of interests is still largely in the experimental stage. It is in the measurement of subjective interests that there has been developed a technique which is as well-defined as in the Binet-Simon Scale of 1908 and more complicated. It may be that a second decade will establish for the field of interests an objective measure. If so, the subjective measures will, of course, go into the discard, unless they contribute something additional to the measurement.

Abilities have always been predominately objective in their definition, even prior to their measurement, while the early conceptions of interests are all subjective. In the field of interests there is the additional problem of finding out exactly what are objective interests, or what is the objective aspect of what are known as subjective

¹⁰ Fryer, Douglas, *The Measurement of Interests*, Henry Holt and Company, New York, 1931, p. 326.

interests, so that the measurement of interests may become an objective problem.

In the statistical treatment of the measuring scales, however, the field of interests is far ahead of the field of abilities of 1908. We know better, today, how to find out the value of an inventory or test than we did twenty years ago. The application of statistical methods to the measurement of interests, particularly to the subjective measures, shows that the scales and scoring keys are of suggestive value—not as suggestive as are measures of abilities today, but fully as suggestive as the ability measures prior to the World War, just before the publication of Yerkes, Bridges, and Hardwick's Point Scale in 1915 and Terman's revision of the Binet-Simon Scale in 1916. The validity coefficients of interest measures, in comparison with those of the measures of abilities today, are promising, and the applications of the interest inventories are as well advanced as the applications of ability scales in 1915.

PART IV
MEASUREMENT OF ACHIEVEMENT

CHAPTER XII

The Measurement of Achievement in Schools

I. What Are Tests of Achievement?

AN ACHIEVEMENT test, theoretically, measures an accomplishment. It differs from an intelligence test, such as we have just considered, in that it measures the achievement without reference to the capacity behind the achievement, whereas the intelligence test aims at measuring the capacity. As contrasted with the intelligence test, which measures primarily native capacity or ability, the achievement test measures primarily something that has been learned or acquired. It is unfortunately rather confusing for the student of school tests that our intelligence tests are so often indirect measures that utilize, as the medium of showing one's mental ability, fundamental material which has been acquired in school. But the intelligence tests do not aim at finding out how much the educational accomplishment is; in fact we utilize as the medium of intelligence testing only that educational content which is rather far behind the individuals being tested and which can be assumed to be more or less constant for all being measured by the test. For example, we do not use reading material to measure the intelligence of the first-graders; we use pictures or some non-verbal material. But for high school students,

208 Measurement of School Achievement

all of whom we can assume have had a fundamental training in the mechanics of reading, we can use a verbal medium for measuring their mental powers.

The need for reliable devices or instruments for measuring achievement arises in many problems involved in dealing with human beings. Throughout the educational world, pupils' achievements in the various school subjects must constantly be measured for purposes of motivating learning, promoting to higher classes or schools, assigning honors and credits, and guiding educationally and vocationally. In the factory or shop, achievements in the form of mechanical skills must often be measured. In the office, achievement measures are in terms of clerical efficiency or typing ability or stenographic ability. Teachers are year after year being measured in some way to indicate teaching efficiency. For far too few of these and other needs for measuring instruments for achievements have reliable devices been worked out. There are numerous achievement tests for school subjects, many of which are very satisfactory; there are a few achievement tests of an informational type for knowledge of a vocational sort; there are a few standardized trade tests for more mechanical accomplishments or skills; and there are attempts at working out rating scales for measuring various total job efficiencies, most of which have not proved reliable in use. This chapter will be concerned chiefly with the first of these four groups; the second and third are patterned largely after the first; and the last—the rating scales—will be considered with our later discussions.

II. Psychology's Contribution to Achievement Testing

Psychology can be given much of the credit for demonstrating the unreliability of the older methods of measur-

ing achievement. It has pointed the way to the development of improved instruments of measurement. As early as 1912 and 1913, Starch and Elliott¹ published articles in the *School Review* on reliability of grading in various high school subjects, in which they demonstrated that with the ordinary type of academic examinations then in vogue, achievement measurements could not be relied upon to be accurate. In one of these studies, the authors selected a final examination paper in geometry written by a high school student. A reproduction of this paper, with the set of examination questions, was sent to 180 high schools in the North Central Association, with the request that it be graded according to the practice and standards of the school by the principal teacher of mathematics. In 116 replies obtained, two assigned a mark above 90 and one below 30. Forty-seven who graded the paper assigned a mark of passing (75) or over; 69, below passing.

More recent investigations of similar measuring instruments corroborate the findings of the earlier studies. Tiegs² reports a study of the grading of a seventh-grade physiology test of five questions by a group of teachers from Los Angeles schools, who were assembled for the purpose of studying tests and measurements. Thirty-one teachers assigned to the paper percentage marks ranging from 20 to 90, distributed as shown in Table XX. As Tiegs states: "These results would indicate that the failure or success of this boy depended upon the teachers

¹ Starch, D., and Elliott, E. C., "Reliability of Grading High School Work in Mathematics," *School Review*, 21: p. 254, April 1913; "Reliability of Grading High School Work in English," *School Review*, 20: p. 442, Sept. 1912; "Reliability of Grading High School Work in History," *School Review*, 21: p. 676, Dec. 1913.

² Tiegs, Ernest W., *Tests and Measurements for Teachers*, Houghton Mifflin Co., Boston, 1931, p. 21.

210 Measurement of School Achievement

fate gave him, rather than upon his knowledge of physiology or his answers to the questions asked."

TABLE XX

PER CENT MARKS ASSIGNED TO PHYSIOLOGY EXAMINATION

<i>Per Cent Marks</i>	<i>Frequency</i>
90-100	4
80-89	8
70-79	3
60-69	2
50-59	9
40-49	4
30-39	0
20-29	1
	<hr/> 31

Such studies as these are all indictments against written examinations of the traditional or essay type. They violate the first rule for a good measuring stick—that it give the same results every time applied for measuring the same thing, no matter who applies it. And yet for decades they have been used as measuring instruments, often with a mistaken confidence in their exactness and accuracy of measurement.

Whatever pedagogical or other values we may justly assign to this type of examination, its shortcomings disqualify it as a measuring instrument. The same criticisms and the same disqualifications apply also to many other schemes for measuring achievements outside the classroom, as the schemes for rating efficiency of teachers and workers.

On a second score, psychology deserves the chief credit for the working out of new methods of measuring achievement. In the classroom these center around the development of "short-answer" type examination questions.

In achievement testing, these were first utilized in the construction of rather formal tests in the specific school subjects, which were standardized, printed, and supplied by test distributing centers for survey testing and other formal programs of measurement. The newly developed intelligence tests furnished the pattern for the types of questions used in the achievement tests. Close upon the introduction of the formal standardized achievement tests came the appreciation of the merits of objective types of examination questions for informal classroom use, and now in many schools much of day-by-day or week-by-week measurement is done by objective tests constructed by the instructor or teacher. As educational workers become more trained in the principles and technique of measurement, the use of these informal objective achievement tests will probably increase.

We have already met many of the objective "short-answer" type questions in our previous considerations. For completeness in our present discussion, the chief types are listed and defined. For actual sample questions, the reader is referred to a subsequent part of this chapter.

1. The true-false question. This type of question consists of a true or a false statement, the truth or falsity of which is to be indicated in some way by the person answering the question or item. Particular advantages of this type of question include relative ease of construction, suitability for a wide variety of material, and quickness with which it can be answered, making it possible to cover a wide field in a short testing time.

2. The multiple-choice question. This type presents to the testee a question or a problem with several (usually three to five) suggested answers or solutions, only one of which is correct. The testee answers the question

by an indication of his choice among the answers. This type of question has found particular use when it is desired to test reasoning and judgment rather than simple information, although the test form can be used for testing the latter also. Contrasted with the true-false form of question, the multiple-choice question is somewhat more difficult to construct and is more time-consuming from the standpoint of answering.

3. The completion question. This form of test question usually requires a one-word or at most a short-phrase response from the testee. It may be in the form of a sentence with an omitted word; or it may be in the form of a question which can be answered by a word, or at most a few words. This type of question is in contrast to the two forms that have already been mentioned, in that it requires recall on the part of the testee, rather than simply recognition.

4. The matching test. This type of test usually consists of two columns of items, each item in one column to be matched with an appropriate item in the other column. For example, a column of dates may be matched with a list of historical events; or a column of authors' names may be matched with a list of books. This type of test has a limited usefulness as compared with other types discussed, because of its unsuitability for many kinds of material on which tests are to be constructed.

In addition to these forms, which meet most of the needs of testing, the student will come across other types of short-answer questions in special situations. Among them may be mentioned (1) identification questions, where parts of drawings or photographs are to be indicated; (2) questions which involve the detection and

correction of errors; (3) classification questions, in which various items are to be classified under a given code or classification scheme; and (4) arrangement in correct order of misarranged events of history or steps in a procedure.

III. Examples of Achievement Tests for Schools

There have been literally hundreds of tests constructed and standardized for the various school subjects. We shall examine some typical ones. No claim is made that the examples selected are the best in the field; they are, however, typical in nature of the material, have been well standardized, and have generally been widely used.

1. **An elementary school reading test.** *The Haggerty Reading Examinations* (Sigma 1 for primary grades and Sigma 3 for upper grades) are among the best known and most widely used tests in elementary school achievement. The primary test consists of two parts, the first in the nature of a directions test involving both words and pictures, and the second a true-false test calling for interpretation of simple sentences. The whole test includes 45 items, requiring about a half hour for answering. The test is thoroughly objective in scoring and has a high reliability. The author reports a reliability coefficient of .84 obtained by correlating the results of two testings on a group of 200 children tested six weeks apart. The validity of the examination has been demonstrated by numerous studies which show that it agrees well with other reliable estimates or measurements of scholastic ability in reading. Age and grade standards are available for interpretation of scores made on the test. A few samples are given from the test to show its nature.

HAGGERTY READING EXAMINATION—SIGMA 1³

PART 1



6. Make two lines under the big bubble that is in the air.

7. Put a cross above the pipe in the girl's hand.

PART 2

1. Can you eat? NO YES
2. Can a hat walk? NO YES
3. Can a clock talk? NO YES
11. Does flour come from milk? NO YES
12. Is every man a soldier? NO YES
13. Are dresses sometimes made of gingham? . NO YES

2. A high school literature test. The George Washington University Series of standardized tests contains a comprehensive test of high school literature by Omwake, Schwarz, and Ronning. This test is divided into three parts: 50 questions in multiple-choice form, 90 in true-false form, and 20 in matching form. The authors state that construction of the test was based upon an analysis of content of English literature courses as stated in catalogues and bulletins of 22 representative states and 15 larger cities of the United States. A reliability coefficient of .90 is reported, based upon retest of 98 high school seniors. Validity of the test is indicated by close agreement of test scores with other objective tests in English and with school grades in literature courses, and

³ Copyright by the World Book Company, Yonkers-on-Hudson, New York. Reprinted by written permission of the publishers.

by the increase in scores from year to year in the high school with increasing study of literature.

ENGLISH LITERATURE TEST ⁴

PART 1

Directions. Write the *number* of the best answer on the line at the right.

1. Lady Macbeth sent the guests away from the banquet because (1) she was frightened by the ghost of Banquo (2) Macbeth was very ill (3) she feared for the safety of her guests (4) she wanted to prepare for the murder (5) she feared Macbeth would betray his crime -----
2. Francis Bacon's chief contribution to literature was in the form of (1) drama (2) essays (3) poetry (4) criticism (5) letters -----
3. Burke considered his proposed plan for dealing with the colonies (1) philanthropic (2) impracticable (3) expedient (4) revolutionary (5) temporary -----
4. Poe's short stories abound in (1) reverence (2) horror (3) humor (4) patriotism (5) romance -----
5. The poetry of Burns is noted chiefly for (1) its matchless English style (2) its accurate pictures of society life (3) its classical allusions (4) its faultless form (5) its sympathetic treatment of homely, every-day themes -----
6. "Robinson Crusoe" is noted for its (1) adventure (2) satire (3) mystery (4) hidden meanings (5) elaborate style .. -----
7. Samuel Pepys is remembered chiefly because of his (1) plays (2) letters (3) diary (4) essays (5) poetry .. -----

⁴ Quoted by permission of Center for Psychological Service, Washington, D. C.

216 Measurement of School Achievement

8. The Black Knight is a character in (1) A Tale of Two Cities (2) The Last Days of Pompeii (3) The Black Cat (4) Ivanhoe (5) The Last of the Mohicans -----
9. "Sohrab and Rustum" is a (1) satire (2) drama (3) comedy (4) pastoral (5) narrative poem -----
10. The early authority on fishing was (1) Isaak Walton (2) Thomas Huxley (3) Samuel Pepys (4) Daniel Defoe (5) Henry Thoreau -----

PART 2

Directions. If the statement is true, encircle the *T*; if it is false, encircle the *F*.

- T F 1. "Il Penseroso" describes the charms of a merry social life.
- T F 2. The hero of "Gulliver's Travels" visited a country inhabited by men six inches tall.
- T F 3. Macbeth desired the death of Banquo because he was jealous of the military achievements of Banquo.
- T F 4. "Pilgrim's Progress" is one of the greatest prose allegories in literature.
- T F 5. Most of Bret Harte's stories deal with high society life.
- T F 6. Samuel Johnson is noted especially for his lyric poetry.
- T F 7. In Shakespeare's time women's parts were played by boys.
- T F 8. In his poem "The Bells," Poe described the process of making bells.
- T F 9. George Bernard Shaw is a well-known dramatist of modern times.
- T F 10. Caliban in "The Tempest" was a kindly spirit.

PART 3

Directions. Place on the line preceding each writer in Section B the number of the writer in Section A of whom he is a contemporary. For example, the number 5 is placed on the line preceding Boswell because he is the contemporary of Johnson, whose number in Section A is 5.

<i>Section A</i>		<i>Section B</i>
1. Chaucer	___5___	Boswell, James
2. Shakespeare	-----	Addison, Joseph
3. Milton	-----	Bacon, Francis
4. Pope	-----	Barrie, James
5. Johnson	-----	Browning, Robert
6. Wordsworth	-----	Burke, Edmund
7. Tennyson	-----	Coleridge, Samuel
8. Kipling	-----	Defoe, Daniel
		etc.

3. **A college chemistry test.** We may select another achievement test from the George Washington University series—one which is a good illustration of an objective test in the college field. The one which has been selected for our example is a test in general college chemistry. This particular chemistry test was used in an extensive study of measurement of teaching efficiency which we shall have occasion to refer to again later on in this book.⁵ The test is well standardized, norms being given for students who have had a year of college chemistry both with and without a background of high school chemistry. The 150 points of the test are distributed in five different parts, which are indicated by the sample questions below:

⁵ See Chapter XIV.

GENERAL CHEMISTRY TEST ⁶

PART 1

Directions. In the statements below, certain words and numbers have been replaced by lettered blanks. Insert the omitted words or numbers in the corresponding blanks at the right.

I. Avogadro's hypothesis states that equal volumes of gases at the same temperature and pressure contain an equal number of _____ *a* (a) _____

II. The gram-molecular weight of a gas at standard pressure and temperature occupies a volume of _____ *b* (b) _____ liters.

PART 2

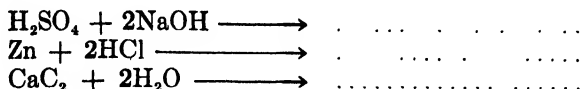
Directions. If the statement is true, encircle the *T*; if it is false, encircle the *F*.

- T F 1. Mendeleef outlined the periodic relations of the elements.
- T F 2. Hydrogen is generally prepared in the laboratory by the reaction of an acid and a metal.
- T F 3. An element always passes from one form of combination to another without change of valence.
- T F 4. Carbon dioxide is made in the laboratory by the reaction of hydrochloric acid and calcium carbonate.
- T F 5. Limestone is an important original source of calcium.

⁶ Quoted by permission of Center for Psychological Service, Washington, D. C.

PART 4

Directions. Complete and balance the following equations.



PART 5

Directions. Solve the following problems. Place the answer on the line at the right. Use the left margin for any figuring needed.

1. One volume of oxygen combines with two volumes of hydrogen to form how many volumes of H_2O in gaseous form? -----
2. What volume of 0.5 normal solution of HCl would be required to neutralize 1500 cc. of 2.0 normal NaOH solution? -----

4. Objective measurement of school achievement by product scales. Certain school subjects which must be graded or rated as frequently as any other subjects do not lend themselves to the type of measurement that we have just illustrated by several sample tests. Such subjects are handwriting, composition, drawing, sewing, and other less frequently rated ones of a similar nature. Tests cannot be composed for these performances or skills in terms of so many questions to be answered, the answers to which can be easily scored as right or wrong. Merit in such subjects clearly depends upon quality of a total performance, and an achievement grade or rating should be indicative of the degree of this quality. Quality judgments left to the free decision of the teacher or rater usually have all the disadvantages of ratings derived from grading essay-type examination questions. Hence scales for the more accurate grading

220 Measurement of School Achievement

of subjects like handwriting and composition have been developed. Some of these scales, in fact, antedate most of the test work that has been done in terms of short-answer type test material. Thorndike's Handwriting Scale appeared in one edition as early as 1909, and Ayres' Scale as early as 1912.

Generally the various product scales consist of several samples of performance which have been carefully selected, graded, and scaled with a range in quality from very poor to very good. Thus, the Ayres Measuring Scale for Handwriting consists of eight samples of handwriting graded from 20 to 90. The Willing Composition Scale consists of eight samples of compositions graded also from 20 to 90. The Murdock Sewing Scale contains photographs in three views of fifteen sewing samples. Students' performances in these various subjects are graded by comparison with the scale samples. The use of the scales insures greater objectivity of rating; establishes greater uniformity of ratings given at various times; and establishes standards for grading that can be compared from teacher to teacher and from school to school. An illustration from one of these scales is given in Fig. 16.⁷

5. Comprehensive achievement tests and classification tests. Should one desire to measure the whole educational achievement over a considerable period of school training, there are several standardized achievement tests covering the whole high school or elementary school field. These are sometimes referred to as classification tests because of their usefulness in classifying pupils into sections on the basis of their educational achievement and intelligence.

⁷ Reproduced by permission of the Russell Sage Foundation, New York, N. Y.

20	50	90
<p>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and dedicated, can long survive. We are testing the proposition that all men are equal.</p>	<p>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and dedicated, can long survive. We are testing the proposition that all men are equal.</p>	<p>Four score and seven years ago our fathers brought forth upon this continent a new nation, conceived in liberty, and dedicated to the proposition that all men are equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and dedicated, can long survive. We are testing the proposition that all men are equal.</p>

Fig. 16.—Samples from the Ayres Handwriting Scale.
(Script of original is blue.)

222 Measurement of School Achievement

In the high school field the *Iowa High School Content Examination*, devised by Ruch, and the *Sones-Harry High School Achievement Test* are typical. Both of these tests contain four parts covering, respectively, English and literature, mathematics, science, and the social sciences.

Probably the most used of the general achievement tests in the elementary school field is the *Stanford Achievement Test*, devised by Kelley, Ruch, and Terman. This is published as a Primary Examination and an Advanced Examination. The Advanced Examination contains parts on the following subjects: Reading—Paragraph Meaning; Reading—Word Meaning; Dictation Exercise; Language Usage; Literature; History and Civics; Geography; Physiology and Hygiene; Arithmetic Reasoning; and Arithmetic Computation. As a test of achievement the Stanford Examination has many advantages. It is published in several forms; its reliability and validity have been established by thorough studies; and reliable norms are available for both age and grade. A feature of this test as well as of some other educational achievement tests is the possibility of deriving from the test standards a measure of educational attainment commonly designated as "Educational Age." For example, if a pupil's total score on the test is equivalent to the average or standard for age 10 years 6 months, we may say that his *Educational Age* is 10 years 6 months. This gives a measure of educational achievement which has more meaning than simply the raw test score. The use of Educational Ages also permits the calculation of *Educational Quotients* by dividing the pupil's Educational Age by his chronological age. Such quotients are similar in meaning to intelligence quotients based on intelligence tests using mental age standards. They give us in one

measurement a relative rating of the child's educational attainments; or, in other words, they show us in one quantitative rating whether or not the child is up to the educational expectations for his age. If he is, his Educational Quotient will be 1.00 or more; if he is not, it will be below 1.00. *Accomplishment Quotients* are also sometimes worked out by a combined use of educational and intelligence tests. Mathematically, Accomplishment Quotient equals Educational Quotient divided by Intelligence Quotient. In general such a measure indicates whether the pupil is attaining in educational accomplishment what he is capable of in accordance with his intelligence. If he is, his Accomplishment Quotient should be above the 1.00 line; if he is not, it should be below.

IV. The Nature of Short-Answer Achievement Tests

In closing this discussion of achievement tests, let us briefly consider the advantages which have given rise to the widespread use of the short-answer questions. Almost all the advantages are either directly or indirectly related to the *objectivity* of the questions. They are said to possess this quality because they can be scored or graded without the play of opinion or subjective judgment on the part of the scorer or grader. Each question is definitely right or wrong and there is not the necessity of making a decision as to the relative merits of each answer given.

Other advantages which are frequently pointed out for "short-answer" achievement tests include the following:

(a) *Short-answer tests are reliable.* They give the same or very nearly the same results when applied at different times to the same group. They give the same results when administered by different examiners or teachers. They give the same results when scored or

224 Measurement of School Achievement

graded at different times by the same grader. And, finally, results on the tests are the same when scoring or grading is done by different teachers or graders.

(b) *Short-answer tests make possible the construction of equivalent examinations.* Many occasions present themselves in the educational world in which it would be highly desirable, we might think even necessary, to have several examinations of equivalent difficulty. Such occasions arise in the comparisons made from year to year in schools and classes; in the administration of state examinations; and in the giving of examinations for such things as college entrance. We cannot adequately compare a sixth grade this year with a sixth grade next year unless they take equivalent examinations or examinations whose difficulties are definitely known. We cannot be fair if one year we give a state examination which fails only 10 per cent of the testees and next year one which fails 30 per cent of an equal-ability group. Ben Wood⁸ some time ago made a study of examinations (traditional type) given by the College Entrance Examination Board. He says:

Large and unrecognized differences were found between the difficulties of the examinations which were thought to be equal. According to the reports of the Secretary of the College Entrance Examination Board, the percentages of failures in algebra for the years 1916 to 1921 inclusive have been as follows:

	YEAR					
	1916	1917	1918	1919	1920	1921
Percentage of failures	61.8	36.7	25.3	61.3	26.1	28.5

These variations are almost certainly due to differences in the difficulties of the examinations used.

⁸ Wood, B. D., "Measurement of College Work," *Educational Administration and Supervision*, 7: p. 301, Sept. 1921.

Because of the subjectiveness of the traditional type of examination question, it would be impossible even with prior statistical study of the questions to construct equivalent examinations. However, when the need arises, a minimum of preliminary analysis of objective question material will insure fairly equivalent examination material in various tests.

(c) *Definite standards can be set on objective tests.* These may be averages of attainment for various grades, for various ages, for groups taught by different methods, for those pursuing different courses of study, etc. Such standards, if set, mean little in traditional essay-type examinations, because if a class fails to meet a stated standard, we are not sure whether the failure represents a lack of attainment among the pupils or too strict grading on the part of the teacher of the class.

(d) *Short-answer type of tests are comprehensive in nature.* Because of the quickness with which the questions can be answered, and therefore the number that can be asked, they are likely to represent a much better sampling of the field being tested.

(e) *Short-answer tests readily show pupils where they are wrong.* For this reason they may be said to possess considerable pedagogical value, though their critics have often praised the traditional type of examination for its greater pedagogical value as compared with the short-answer examination. It is true that short-answer examinations do not show where the reasoning went wrong in incorrect responses, but this can hardly be claimed in most instances for any other type of examination.

(f) *Short-answer tests are easily scored.* Many of them do not even require a knowledge of the subject on the part of the scorer. A clerically accurate person is an adequate scorer in many instances.

226 Measurement of School Achievement

(g) *Pupils and other testees as a rule like short-answer tests.* The tests appeal because of their ease of taking, because of their wide sampling of the field, and because of their fairness.

CHAPTER XIII

Objective Achievement Tests in Professional Schools

THE need for objective test methods in measuring achievement in the higher levels of education is just as great as in the lower schools. Universities and professional schools must utilize measurements of the achievements of their students for assigning class grades; for advising students educationally and vocationally; for recommending them for occupational and vocational placements; and, frequently, in the professional schools at least, for giving them initial certifications to enter the various professions, as medicine, law, etc.

I. The Difficulties Encountered

The development of objective tests, particularly well-studied standardized tests at the university and professional school level, has met with certain difficulties which we may consider greater than those at the public school, or elementary and high school, level. In the first place, the material of many university courses is less standardized in content from course to course or college to college. This fact limits the usefulness of standardized tests, but it does not affect the usefulness of less formal examinations constructed on objective test principles. In the second place, the nature of the subject matter of the courses often makes the construction of the objective-test type of questions more difficult than for other edu-

cational levels. This difficulty, however, is not insurmountable, and can usually be overcome by training and practice on the part of the test maker in the preparation of objective test material. In the third place, university instructors have often hesitated to accept the objective type of test for measurement of their students because they believe that the short-answer type of question fails to measure qualities in keeping with important aims of the course. To these instructors, the short-answer objective test does not measure, to a sufficient degree, qualities such as reasoning, organization of material, and judgment about problems of the course. This criticism may be justified in the case of poorly constructed tests, but it is hardly a fair criticism of well-made ones—good tests of the objective type can measure all these qualities. And even those objective test forms which do measure primarily memory and purely informational knowledge about a course correlate highly with test forms which aim primarily at measuring reasoning, judgment, or organizational ability as applied to the same subject matter.

II. Early Studies in Achievement Testing in Professional Schools

The application of objective test principles in the higher professional schools began in the early twenties. Ben Wood's book *Measurement in Higher Education* contains an examination of this type administered in mechanical engineering courses in 1922. During the school year of 1924-5 a study of objective test methods in medical school subjects was undertaken in certain departments of the George Washington University Medical School by Moss and Hunter.¹ The following year, sim-

¹ Hunter, Oscar B., and Moss, F. A., "Standardized Tests in Bacteriology," *Public Personnel Studies*, Vol. III, No. 2, February 1925, p. 52.

ilar studies were made of objective tests in certain subjects at the College of Physicians and Surgeons, Columbia University, under the guidance of Wood.² The foreword to Wood's report, written by the Dean of the College of Physicians and Surgeons, expresses a favorable and hopeful attitude toward the use of objective measurements of achievement in the professional school:

Anyone who follows the work of a class of students in medicine is surprised each year at the showing made by a considerable proportion of the group. The marks given their examination papers are often surprisingly high and quite as often surprisingly low. This leads the one who corrects the old fixed type of examination paper to wonder as to his ability as a critic or as to the dependability of the student, or perhaps, of the examination as a test of his mental equipment.

The idea that an examination in the Medical School could be prepared so as to give a more exact estimate of the student's ability was received by the members of the staff with enthusiasm and real hope, even if associated with some doubt. That such a plan would also decrease the labor connected with rating and grading such examinations brought even greater joy. Therefore, when Professor Wood discussed the project he had in mind, it was taken up by several of the departments with enthusiasm.

We are trying to be conservative and reserve our final decision in the matter, but the results of this, his first study with us, make us at least optimistic for the future. We sincerely appreciate his efforts and trust that the studies may continue.

III. An Objective Bacteriology Test

As an example of measurement of achievement at the professional school level, we shall discuss one of the tests

² Wood, Ben D., "New Type Examinations in the College of Physicians and Surgeons," *The Journal of Personnel Research*, Vol. V, No. 6, October 1926, p. 227.

developed in the George Washington study—a test in Bacteriology. This test consists of four parts, as follows:

Test 1—Organisms, Methods, Infection, and Immunity: Section A of this test consists of 50 questions in the multiple-choice form, and Section B of 50 questions in the true-false form. *Test 2—Recognition and Diagnosis of Bacterial Micro-Organisms:* In Section One are the names of 15 micro-organisms, 10 of which are described in Section Two. The person being tested writes the name of the organism listed in Section One which most nearly fits the description given in Section Two. *Test 3—Laboratory Procedure:* Four procedures, each of which can be accomplished in five steps, are listed, but not in the order in which they would actually be done; in addition, several useless steps are included. For each procedure, the person taking the test numbers from one to five the steps that he would use to carry out the procedure, in the order in which he would use them. *Test 4—Identification of Micro-Organisms from Lantern Slides:* Ten lantern slides of pathogenic bacteria, showing typical colonies, various cultural characteristics, cell groupings, and the typical morphology of the organisms, are shown. From the slide the organism must be identified. Portions of this test are reproduced below as sample material.³

TEST I—ORGANISMS, METHODS, INFECTION, AND IMMUNITY

SECTION A

Directions. For each of the following questions four answers are suggested. Before each answer is a space in which to make a mark. Read over the four answers and then place a cross (X) in the space before the

³ Quoted by permission.

answer which is best or most nearly correct. Do not place a cross (X) before more than one answer under each question; if you do, your work on that question will not be counted. Only one of the four answers is the best one and you are to show by the cross (X) which it is.

1. Where the presence of an organism reinforces or augments the growth of another organism that condition is spoken of as:
 - Saprophytism
 - Parasitism
 - Symbiosis
 - Plasmolysis
2. It has been demonstrated that the disinfectant power of most chemical substances is proportional to:
 - Concentration
 - Dissociation into ions
 - Solubility in alcohol
 - Temperature at which boiling takes place
3. The one of the following micro-organisms which usually occurs intra-cellularly is the:
 - Staphylococcus aureus
 - Bacillus diphtheriae
 - Gonococcus
 - Bacillus tuberculosis
4. In sterilizing in the hot air chamber the heat is kept for about one hour's time at a temperature of:
 - 50 degrees C
 - 100 degrees C
 - 150 degrees C
 - 200 degrees C
5. Material rich in albuminous substances is usually sterilized:
 - In a hot air chamber
 - In an Arnold

Objective Achievement Tests

- In an autoclave
 - In an inspissator
6. A substance which inhibits growth and multiplication of bacteria, but does not destroy them, is called:
- A germicide
 - A disinfectant
 - An antiseptic
 - A sterilizer
7. For substances not injured by high temperatures or moisture, the quickest method of sterilization is by:
- The inspissator
 - The Arnold sterilizer
 - Boiling
 - The autoclave
8. The "Arnold" is often used in sterilization by
- The hot air method
 - Steam under pressure
 - Exposure to live steam at 100 degrees C
 - Boiling
9. Facultative anaerobes are bacteria which:
- Cannot grow in the presence of free oxygen
 - Cannot grow without free oxygen
 - Prefer free oxygen but can grow without it
 - Prefer an environment in which there is no free oxygen but can grow in the presence of free oxygen

SECTION B

Directions. Examine each statement below and decide whether it is true or false. If the statement is true, encircle the *T*. If it is false, encircle the *F*.

- SAMPLE: (T) F The streptococci occur in chains.
T (F) The chief cause of tuberculosis is the typhoid bacillus.
- T F 1. Organisms that retain the gentian-violet are said to be Gram positive.
- T F 2. Plasmoptysis occurs when the cell is removed from a medium of high osmotic pressure to one of low osmotic pressure.
- T F 3. The development of bacteria is arrested more often by accumulation of waste products than by exhaustion of nutrient material.
- T F 4. The term immunity as used by bacteriologists means the inability of an animal to become infected by a certain micro-organism.
- T F 5. The characteristic affinity of specific bacterial poisons for certain tissues is generally recognized by bacteriologists.
- T F 6. As a rule the mutations produced in a micro-organism readily revert to type when the micro-organism is subjected to the proper environment.
- T F 7. Sporulation is the most common method of multiplication in many forms of bacteria.
- T F 8. All bacteria require oxygen in some form to grow and reproduce.
- T F 9. Low temperatures are on the whole more destructive of bacteria than high ones.
- T F 10. In the case of the halogens the germicidal power is directly proportional to their atomic weight.

TEST II—RECOGNITION AND DIAGNOSIS OF BACTERIAL MICRO-ORGANISMS

Directions. In Section One below are the names of fifteen organisms, ten of which are described in Section Two. Read each of the ten descriptions and write on

the line after it the name of the organism in Section One which most nearly fits the description.

SECTION ONE

- | | |
|---------------------------------|---------------------------------|
| 1. <i>Pneumococcus</i> | 8. <i>Streptococcus</i> |
| 2. <i>Bacillus anthracis</i> | hemolyticus |
| 3. <i>Bacillus tetani</i> | 9. <i>Staphylococcus aureus</i> |
| 4. <i>Bacillus diphtheriae</i> | 10. <i>Gonococcus</i> |
| 5. <i>Bacillus subtilis</i> | 11. <i>Bacillus pestis</i> |
| 6. <i>Bacillus tuberculosis</i> | 12. <i>Bacillus pyocyaneus</i> |
| 7. <i>Bacillus coli</i> | 13. <i>Bacillus typhosus</i> |
| communior | 14. <i>Meningococcus</i> |
| | 15. <i>Spirochaeta pallida</i> |

SECTION TWO

1. A non-motile, facultative anaerobic, non-liquefying, non-chromogenic, acid fast, parasitic, non-spore bearing, gymnobacterium, highly pathogenic for man, not readily stained by the usual aniline dyes; Gram positive; growing only on special media, exhibiting many involution forms in old cultures, but usually seen in fresh smears as a delicate, slender, slightly curved rod with rounded ends and beaded body _____
2. A small, coffee-bean shaped, non-motile, facultative anaerobic, non-liquefying, non-chromogenic coccus, spontaneously pathogenic only for man, staining readily with the usual aniline dyes; Gram negative; produces no spores, has no capsules; can be grown only on special media; usually seen intracellularly in smears made directly from fresh pus _____

TEST III—LABORATORY PROCEDURE

This test consists of four procedures frequently performed by a bacteriologist. Each procedure can be accomplished by five steps. The five steps are given, but not in the order in which they would actually be done.

In addition to the five steps actually necessary several useless, additional steps are listed. Look over all the steps suggested for the procedure and number the steps from 1 to 5 that you would use in the order in which you would use them.

SAMPLE: Procedure for making a simple stain;
3 steps necessary:

- 1 Make a thin smear and fix it on a slide by
----- passing through the flame.
- Decolorize with alcohol.
- Apply Gram's iodine solution.
- 3 Wash off excess stain with water.
- 2 Apply methylene blue stain.
- Wash in chloroform for one minute.

1. Procedure for making a Gram stain; 5 steps necessary:

- Put on Gram's iodine solution and wash off
excess stain with water.
- Decolorize with 95 percent alcohol and wash
off excess alcohol with water.
- Wash off excess chloroform with water.
- Apply 5 percent aqueous carbolic acid.
- Put on methylene blue.
- Wash in chloroform for one minute.
- Counterstain with safranin and wash off ex-
cess stain with water.
- Make a thin smear and fix on slide by passing
through the flame.
- Steam over the flame for three minutes.
- Apply aniline gentian violet and wash off ex-
cess stain with water.

TEST IV—IDENTIFICATION OF MICRO-ORGAN-
ISMS FROM LANTERN SLIDES

You will be shown ten slides, each showing the morphology and cultural characteristics of a micro-organism. You will be allowed one minute to study each slide and identify the micro-organisms. After the

slide is removed you are to write the name of the micro-organism in the appropriate space below.

- | | |
|----------|-----------|
| 1. _____ | 6. _____ |
| 2. _____ | 7. _____ |
| 3. _____ | 8. _____ |
| 4. _____ | 9. _____ |
| 5. _____ | 10. _____ |

IV. Study and Standardization of the Tests in Bacteriology

In the validation and standardization of the tests in bacteriology, the first step was to find a considerable group of persons whose abilities could be determined with reliability. The group used in the study of these tests included two classes in bacteriology taught during the school year of 1924-5 (altogether 100 students). The painstaking coöperation of three instructors of these students made possible the establishing of a reliable criterion of their abilities against which to check the new test material.

The criterion used was a rating, established as follows: Each day during the course the class was divided and given a one-hour quiz by the professors; the professors alternated sections each day. Every student was called upon at each quiz and was rated according to his answers. These same professors also observed the men in the actual performance of the laboratory work, and in addition to this they had given the men three written examinations previous to making their ratings. In that way the professors whose ratings were secured were enabled to know their men quite well.

In estimating the abilities of the students the three judges each made independent estimates, using the 7-point scale described below. It was definitely under-

stood by each of the three that the estimates were to be based purely on the man's knowledge of the subject and should in no case be influenced by his personal traits, study habits, or scholarship in other subjects. Each of the three judges agreed to the following standards of rating:

The ratings should be based on an all-around estimate of the man's knowledge of bacteriology. His personality and his habits (such as neatness and persistence) should not be taken into consideration in any way whatever in making up this rating. This judgment should be based solely upon the judge's opinion of the man's information about the subject, as manifested in his daily quizzes, his laboratory work, and his monthly examinations.

The students are to be rated in seven groups. The 4 or 5 very outstanding individuals are to be given a rating of 7, or exceptionally superior. The 9 or 10 individuals who are noticeably above average are to be given a rating of 6. The 18 or 20 individuals who are slightly above average are to be given a rating of 5. The 30 or 35 individuals who are just average, fair, or ordinary are to be given a rating of 4. The 18 or 20 individuals who are slightly below average are to be given a rating of 3. The 9 or 10 individuals who are noticeably below average are to be given a rating of 2. The 4 or 5 individuals who are exceptionally poor are to be given a rating of 1.

After each of the three judges had independently prepared his estimates, the three held a conference to try to discover the reasons for differences. The cases where their estimates differed more than two points on the 7-point scale were brought up for special consideration in order to determine why their agreement was not closer. In practically all these cases one or the other of the judges had either overlooked something that the other

judges knew, knew some trait that the other judges did not know, or was estimating on something that could not properly be considered as part of the student's knowledge and information on the subject of bacteriology. As a result, each of the judges made some revisions in his estimates and in the end there was a much closer agreement than before, the final result being that in only a few instances were there more than two points of divergence between the ratings of any two judges for any individual.

The revised ratings were then added and the average taken. The first 25 of these, with the bacteriology test scores, are given in Table XXI.

TABLE XXI
RECORDS FOR BACTERIOLOGY TEST STUDY

STUDENT'S NUMBER	—INSTRUCTORS' RATINGS—				SCORE ON BACTERIOLOGY TEST
	Judge A	Judge B	Judge C	Average	
1	7	7	7	7	193
2	7	7	7	7	191
3	5	7	7	6.3	173
4	6	6	6	6	176
5	6	6	6	6	175
6	6	6	6	6	173
7	6	6	6	6	172
8	6	6	6	6	172
9	5	6	7	6	171
10	6	6	6	6	166
11	6	6	6	6	166
12	6	7	5	6	164
13	6	7	5	6	163
14	6	6	6	6	163
15	5	5	7	5.7	181
16	6	6	5	5.7	153
17	5	5	6	5.3	163
18	5	5	5	5	168
19	4	5	6	5	165
20	4	5	6	5	164
21	5	6	4	5	162
22	4	5	6	5	160
23	4	6	5	5	155
24	5	5	5	5	150
25	6	5	4	5	143

The correlation coefficient between the criterion of bacteriological knowledge and ability and the total test scores on the new type examination is plus .79, a much higher relationship than we can find between studies that have been made of validity of old type examinations. The reliability coefficient of the new type material is plus .90, as compared with the usually low reliability coefficients for old type examinations.

The study of these tests in bacteriology also included the establishment of tentative norms based upon those cases included in the validity study and additional cases tested in other schools. The norms as stated are given in Table XXII.

TABLE XXII

NORMS ON BACTERIOLOGY TEST

10% made 167 or more	60% made 132 or more
20% made 158 or more	70% made 126 or more
30% made 148 or more	80% made 117 or more
40% made 142 or more	90% made 101 or more
50% made 137 or more	

The authors point out the following advantages of the new type test development: (1) It allows for the possibility of making a wide sampling of information and judgment; (2) The giving and rating of the tests is characterized by ease, exactness, and quickness; (3) The tests allow of the possibility of standardization; (4) Additional sets of tests of equal difficulty can be constructed; (5) The abilities (in bacteriological success) to be measured are wide and varied, so that varied types of tests as measuring instruments seem suitable; (6) Those taking the new type tests found the material interesting and, as a test, fair and to the point.

CHAPTER XIV

The Measurement of Job Efficiencies

SO FAR we have limited our consideration of measurement of achievement to the accomplishments of students in the schools. There is at least one other big group of individuals whose achievements must constantly be weighed in the balance. This group is made up of the employees of industry, business, and government. The significance of an accurate means of measuring their achievements (or efficiencies, to use the employer's term) can hardly be overestimated. Such measurements are the logical bases for promotions and other rewards for good service; they are the starting point for adjustment procedures in cases of unadjusted workers; they determine the order of lay-offs in time of unemployment; they are the criterion against which employment methods are checked.

In a small percentage of cases, job achievements or efficiencies can be quantitatively measured by a simple count of production. The performance of the factory worker who folds handkerchiefs all day can be measured by the number he folds; or the automobile salesman's efficiency may be measured by the number of cars he sells. But the vast majority of workers are not performing duties which can be easily measured. It is for this large group that psychologists, employers, and personnel administrators have combined their efforts in attempting to devise means of accurately measuring job efficiencies.

Where direct quantitative measures of human performances or human traits are not available, and where some measurement seems necessary, resort has invariably been had to some sort of estimation by superiors. These estimates have constituted the "efficiency ratings" or "service ratings" assigned to employees by superiors, foremen, or employers. Much of the contribution of psychology to measurement of job efficiencies has been in the improvement of methods of assigning service ratings.

We shall examine, as a sample, one study in the field of service ratings. (For a discussion of common types of rating scales the reader is referred to Chapter XIX.) We shall also discuss one example of the application of objective tests in arriving at achievement measures in a vocation that has commonly been the source of much trouble because of the subjectiveness and inaccuracy of assigned efficiency ratings.

I. A Service Rating Scale

About 1928, J. B. Probst began work on a service rating scale which might serve as a basis for accurate measurement of job efficiencies in a great variety of jobs. His scale was later extensively studied through a grant to the Bureau of Public Personnel Administration. In the report on this study, Mr. Telford, then director of the Bureau, took a very hopeful attitude toward the possibility of securing reliable measures of job achievements through use of the new rating scale. We quote from his foreword to the report.¹

Personnel administration as yet falls in the category of activities whose success or failure is not closely measurable. The esteem in which a personnel system is

¹ Probst, J. B., *Service Ratings*, Technical Bulletin No. 4, Bureau of Public Personnel Administration, Chicago, p. 5.

held depends only to a small degree upon the extent to which positions are classified with reference to the duties and responsibilities of their incumbents, the absolute and relative levels of pay for various kinds of work, the extent to which a competent or incompetent personnel is secured and retained, and the methods used in fixing hours of work, checking attendance, granting annual and sick leaves of absence, and making demotions and removals. It is hardly too much to say that in most organizations, both large and small, both public and private, the esteem in which the personnel system is held depends not upon whether the bridge carries the load or falls into the stream, but upon whether those things which are undertaken are done with a certain degree of diplomacy so as to avoid giving offense to the persons involved.

The need for a measuring instrument to indicate the degree of success or failure in personnel work has for a good many years been generally recognized. Time after time the attempt has been made to devise, install, and operate a system of service ratings for this very purpose. Invariably, however, the results have been such that the most ardent advocates of these systems have soon perceived their failure actually to measure performance. Neither the management, the administrative and supervisory officers, the employees, nor the central personnel agency have long believed that the so-called efficiency rating system really indicates efficiency or anything else of particular value.

In view of the numerous unsuccessful attempts to devise and operate systems which really measure performance on the job, either relatively or absolutely, *a priori* reasoning would lead to the conclusion that the system evolved by Mr. J. B. Probst, which is described in this book, would be neither much worse nor much better than its predecessors. Extensive experienced work with rigid statistical analyses of the results, as well as a different approach to the subject, has, however, led those connected with the Bureau of Public Personnel Administration and the Civil Service Assembly of the United

States and Canada to the belief that Mr. Probst actually has produced a measuring instrument which, while still not as accurate as those devices which determine the diameter of a bolt to the thousandth of an inch, is fairly comparable to the yard stick graduated into feet and inches.

The service rating study resulted in the development of the Service Report Form and a scoring system for deriving from the superior's reports a quantitative rating for each employee rated. On pages 244 and 245 are reproduced in reduced size the front and reverse sides of the report form. The general nature of the service rating scheme and some of its characteristics as compared with less successful plans of rating may be indicated in terms of the aims set by the author. He states that ² "an attempt was made to develop a system that would eliminate . . . particularly the following:

- (a) "Halo."
- (b) Adjustments of ratings to harmonize high and low raters.
- (c) Inconsistency between reports or ratings made at different times by the same person.
- (d) Need for judging the various employees as to their relative excellence in each trait.
- (e) Necessity for special training of the reporting officer.
- (f) Judgment ratings that admit of little chance for review.

It was decided at the outset that the new system must provide, as far as possible:

- (a) That the employee's performance be reported not merely in general conclusions but in statements of fact or specific and verifiable judgments.

² *Ibid.*, p. 22.

Measurement of Job Efficiencies

- (b) That the facts, traits, or qualities be stated on the report form in terms of the everyday thinking of the reporting officer, not in letters or percentages.
- (c) That the reported facts be properly interpreted and evaluated by a process which, in the very nature of things, must be developed through extensive experiments based on thoroughly sound principles.

THIS FORM IS SPONSORED BY THE BUREAU OF PUBLIC PERSONNEL, AND IS USED EXCLUSIVELY BY THE BUREAU IN THE MEASUREMENT OF EMPLOYEES' EFFICIENCY. SERVICE REPORTS FOR BOTH PUBLIC AND PRIVATE EMPLOYMENT.

THE PROBST SERVICE REPORT

FOR APPRAISING AN EMPLOYEE'S SERVICE VALUE

FOR THE SIX MONTH PERIOD

ENDING _____

THE FACTS AND COMMENTS HEREON SHOULD BE SET OUT AND EVALUATED BY A SUPERVISOR, EMPLOYEE, OR A COMMITTEE OF BOTH. IN A LETTER OF APPRAISAL, THE REPORTING OFFICER SHOULD SET OUT THE FACTS.

RATING

NAME OF EMPLOYEE _____

TITLE _____

DEPARTMENT OR DIVISION _____

INSTRUCTIONS

- On this form you are to report the service value of the employee mentioned above. The report should be for the six-month period shown hereon, unless otherwise indicated.
- In addition to the blanks to be filled in on this side of the sheet, you should check (with an X) all those items on the other side that you can find within it properly fit or describe this employee. Do not guess; if you are not reasonably sure that the employee possesses the trait or quality indicated by a certain item, do not check that item at all. It is not necessary to check any given number of items. You may be able to check 25 or more for one employee and have difficulty in finding more than a dozen or so to describe properly some other employee. Make your X's small, keep them inside the little squares. Do not change the wording of any item.
- This sheet should be checked independently by three supervisory officers, wherever possible. Each officer should select one of the three check columns in which to make his X's, and should keep all his marks within that same column on both sides of the sheet. The officer who is lowest in rank or authority should be the first to check the sheet, then the next higher (or equal) in rank should check, and the one in highest authority should check last.
- Some items, such as "Pleasing and mellow voice," "Active and strong," "Good leadership in emergencies," and a few others, should be considered only if they are deemed essential or desirable for the particular position.

How many days was this employee absent during this period— (Do not include absence for leave or time off or absence for regular vacation)

(a) For sickness, with pay _____ days

(b) For sickness, without pay _____ days

(c) For personal reasons, with or without pay _____ days

(d) How many days suspended, if any _____ days

(e) If there were any fine hours of pay loss of six days or more, or other penalty not resulting in suspension or absence without leave, explain briefly here: _____

Check only one item in each of the following boxes. In doing this, you may consider not only the punctuality of the employee in reporting for work, but also in answering calls, keeping appointments, and handing in reports.

BOX 1

BOX 2

BOX 3

Check Columns			
1	2	3	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Nearly always late
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually late
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often late (about half the time)
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually punctual
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Never, or hardly ever, late
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Nearly always quits ahead of time
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually quits ahead of time
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often quits ahead of time
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Watches clock too much near quitting time
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Seldom quits ahead of time
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Never quits ahead of time

DO NOT CHECK IN THIS BOX FOR THE UNEMPLOYED POLICE OR FIRE PERSONNEL

Probst Service Report (Obverse Side)

- (d) That the report sheet permit the officer an optional selection of various traits and qualities, so that he may report on only those things with respect to which he has definite knowledge.
- (e) That the reporting officer be not required to measure relative degrees of a quality in different employees.

DIRECTIONS: Place an X mark next to each of the items on this page which you know from your own knowledge will describe or fit this employe. Do not guess; check only if you are reasonably certain.

Check Columns				Check Columns			
1	2	3		1	2	3	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Lazy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Active and strong
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Slow moving	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Active but not strong
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Quick and active	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually careless of personal appearance
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Too old for the work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually neat personal appearance
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Minor physical defects	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Pleasing and mellow voice
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Serious physical defects	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Very tactful in dealing with the public
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Indifferent	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Poor technical training for the work
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Talks too much	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good technical training for the work
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Too blunt or outspoken	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Poor head work in sudden emergencies
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Too much self-importance	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good head work in sudden emergencies
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good team worker	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often assigned to other important positions*
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Not a good team worker	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often assigned to fill a higher position*
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Resents criticism or suggestions	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Sometimes goes on a "tear"
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Antagonizes when dealing with others	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Drunk is one of principal failings
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Might often be more considerate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Willing worker, but is not a leader
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually pleasant and cheerful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Does not plan or lay out work effectively
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Always courteous	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Plans work well but lacks snap in getting it done
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Cranky disposition	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Unusual ability in planning and laying out work, good organizer
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often seems dissatisfied	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Makes quick and accurate decisions
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often grumbling or complaining	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Too lenient in maintaining discipline
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Uses poor judgment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Maintains good discipline
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often uses better judgment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Discipline too harsh
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Generally uses good judgment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Exceptionally skillful in handling difficult situations
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Always uses good judgment	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Poor physical appearance
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Does not do his (her) share of work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Has the detective instinct
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Generally looks for the easy work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good knowledge of crooks
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Must generally be told what to do	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Fine memory for faces
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Work often slightly behind	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Lacks courage
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Work always up to date	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Cool headed
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Turns out unusually large amount of work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Has unusual courage
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Steady worker most of the time	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good smoke eater
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Always busy at work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Cannot stand smoke
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Does not accept responsibility	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Poor climber
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Accepts responsibility	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Good climber
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Does not always obey orders willingly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Visits too much with others	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Needs considerable supervision	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Works well without supervision	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Fine self-control, seldom loses temper	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Loses temper easily	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Easily rattled	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Lacks self-confidence	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Too easy-going	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Learns new work slowly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Learns new work easily	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Understands instructions readily	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	A willing worker at all times	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Takes unusual interest in the work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Might be more orderly	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Very orderly and systematic	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often forgetful	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Often does careless work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Makes many mistakes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually accurate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Hardly ever makes a mistake	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Accurate but very deliberate	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Is highly expert in own work	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Not generally reliable or dependable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Usually reliable and dependable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	Always reliable and dependable	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

*Temporary assignments

REMARKS

- (f) That the rating and scoring system make it unnecessary to "adjust" the resulting scores or ratings.
- (g) That the reporting officer be virtually forced to report accurately or be shown by internal evidence in his own reports not to have done so.
- (h) That the results of the scoring system be sufficiently simple for the employees to understand, so that they themselves may easily determine in a general way the fairness and reliability of the distribution of ratings.
- (i) That the scheme also take into consideration the ordinary mental processes of the reporting officer—his reluctance generally to rate negative qualities, his normal desire to say good things about an employee, his tendency to use superlatives in describing the favorite employee, and the like.

The derivation of a rating from the checked report sheet is based on numerical values assigned to the various items in the report. The values are positive for some items and negative for others. The first part of the scoring process is to count the number of favorable, or credit, items checked on the report; the resulting total is termed the *X* score. Then the sum of all the values assigned to all the checked items, favorable or unfavorable, is obtained; that sum constitutes the *Y* score. Numerous experiments were conducted to establish a means of translating these *X* and *Y* scores into reliable and valid ratings, meeting all statistical and technical requirements. Finally a rather complex formula was evolved, and ratings obtained with it have been found to reflect with considerable accuracy the worth of employees rated. The *X* and *Y* scores are first translated into a numerical rating and can subsequently be expressed in letter ratings if desired. The scoring has been reduced to a purely mechanical procedure through the use of special stencils and scale rules. The reader is referred

to the author's report for details regarding the scoring. Probst emphasizes in his report that the officers or supervisors making out the report on an employee are not "rating" him—they are merely indicating certain facts (sometimes approximating judgments) about him—and that the ratings are obtained from the report by a system of evaluating the facts and judgments reported in accordance with a procedure evolved by extensive experimentation. This the author believes to be one of the main advantages of his rating method as compared with those commonly used methods in which the reporting officer attempts to assign relative ratings or values to the various traits or qualities he considers in employees.

General value for a system of rating efficiency of employees can be assumed if (1) the system gives a distribution of ratings approximating the normal distribution curve; (2) the ratings are reliable as shown by agreement between separate ratings made on the same group of employees; and (3) ratings are valid as shown

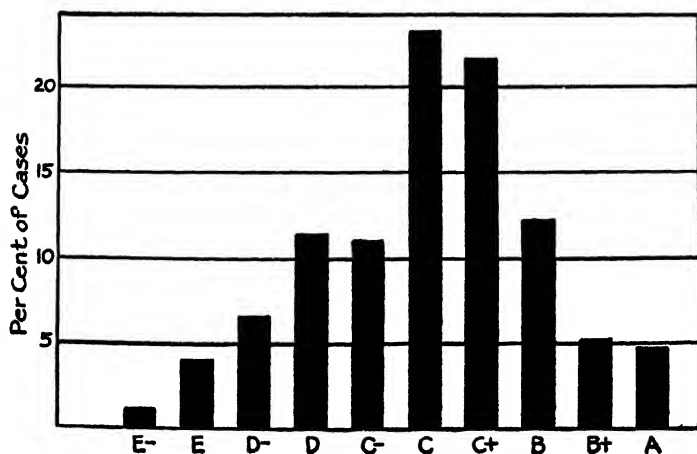


Fig. 17.—Distribution of Probst Service Ratings.

by agreement with reliable criteria of employees' efficiencies.

The distribution of Probst Service Ratings. The distribution of ratings for employees in three cities is shown by the vertical bars in Fig. 17. The form of the distribution closely approximates that of the normal distribution curve. A great many rating scales for evaluating employees' achievements have given very little distribution of the ratings, probably because of the tendency of raters to bunch all the ratings at the high end of the scale. When we know that all the employees in a group are not of equal efficiency, a rating system is

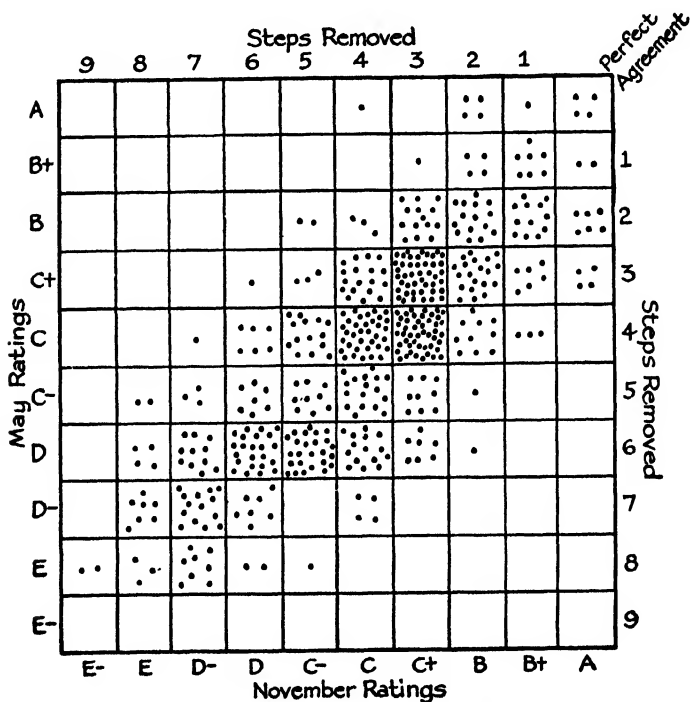


Fig. 18.—Chart Showing Reliability of Probst Service Ratings.

certainly ineffectual if it does not differentiate among those being rated. In view of this common defect of rating systems, it is encouraging to find distributions such as the three illustrated.

The reliability of Probst Service Ratings. Reliability of a rating system is indicated when it produces comparable results in different trials under similar conditions. Probst reports a correlation of .78 between two successive semiannual ratings for a group of 475 municipal employees in the City Water and Power Department of Los Angeles. Fig. 18 shows graphically the relationship. An analysis shows that 155, or 32½ per cent, showed no change in their ratings from one period to the next. Of the remaining number, 217 showed a change of one step, 75 a change of two steps, 26 a change of three steps, and 2 a change of four steps. The changes of one step can be regarded as practically negligible. Such a degree of reliability as shown by the ratings of this group approximates the reliabilities obtained by objective psychological tests, and can certainly be considered high enough to recommend the use of such a rating system as compared with any system of personal estimates unaided by a rating device or system.

The validity of Probst Service Ratings. If the ratings are valid, the employee who is rated *A* and the employee who is rated *E* must in actual fact be *A* and *E* employees, respectively. Such validity is the very essence of the value of the rating system, and yet, unfortunately, it is the most difficult factor to prove. The difficulty rests upon the fact that there are no reliable criteria in most instances of the actual value of the employee with which to compare the ratings. Probst reports numerous studies of validity of the ratings in which Probst Service Ratings have been compared with various

supervisor, foreman, and employer estimates of employee worth. His groups include studies of office workers, laborers, nurses, policemen, firemen, teachers, and others. The large majority of his correlations are above .60, indicating considerable promise for his rating system, especially in the face of a somewhat unknown reliability of the criteria.

The foregoing material has been presented, first, as an example of a rating scale and rating system applicable to measurement of employee achievement; and second, as an example of about the most that can be expected from rating systems in their present stage of development. The reader is warned that with less carefully worked out rating systems, with poorer coöperation on the part of rating officers and supervisors, and perhaps in some occupations, results comparable with those obtained by Probst cannot be expected.

II. Objective Psychological Tests in the Measurement of Job Efficiencies

For a few specific jobs, objective "psychological" type efficiency tests have been devised and standardized. There are, for example, standardized typewriting tests, shorthand tests, bookkeeping tests, teaching efficiency tests, and trade tests in a few of the mechanical trades. These presumably can be administered at periodic intervals to employees on the job and be used as a basis for evaluating their job efficiency. Besides being available for only a very limited number of positions, they often fall short of satisfactory measurements in two respects: they often do not test all the aspects of job success; and they often represent artificial test situations somewhat removed in nature from the actual job itself. However,

supplemented by other considerations and measurements, these tests are valuable tools.

An interesting use of "psychological" tests has frequently been suggested and in a few instances carried out as a basis of measuring teaching efficiency. In 1928-29, the writer coöperated in such a study of teaching efficiency. Let us look very briefly into the nature of this study.

The study took as its starting point the premise that the important purpose of teaching is to make desirable changes in the students, these changes being indicated in the amount the students learn. The purpose of teaching being thus defined, the teaching efficiency of any instructor may be considered to vary directly with the desirable changes produced in the students by his teaching. To arrive at a fair measure of student improvement, it would seem necessary, then, to know three things: the ability or natural capacity of the students to learn; what they know about the subject before the instruction; and what they know at the end of the instruction.

Twenty colleges agreed to coöperate in an experimental rating, on such a basis, of efficiency in teaching general college chemistry. The same standardized test covering general college chemistry was given at the end of the year's work to all the students of the instructors included in the study (6,667 students). Could it have been assumed that all the instructors were teaching students of the same ability and the same chemistry knowledge and training prior to their taking the course, the average test performances of their students after a year's instruction might be considered indicative of their relative teaching efficiency. Since equalities in ability and previous training could not be assumed for the various classes, these

two factors were studied and the raw chemistry-test averages were corrected, where necessary, to equalize ability and previous chemistry knowledge of students of the different instructors. This was done by an adjustment of scores for those students who had already studied chemistry in high school in accordance with the amount of increase in final score such previous training was found to produce. Corrections for differences in average ability of different classes of students were made on the basis of statistical analysis, which showed that a change of one point in intelligence (as measured by the test used) affected the chemistry score by .32; so that, if an instructor's class was 1.0 point above the average for all in-

TABLE XXIII
COMPARISON OF EFFICIENCY OF INSTRUCTORS^a

<i>Instructor</i>	<i>Degree of Instructor</i>	<i>Rank of Instructor</i>	<i>Years' Experience</i>	<i>Teaching Load (Hrs. per week)</i>	<i>Intelligence Median of Class</i>	<i>Chemistry Median (Uncorrected)</i>	<i>Chemistry Median (Corrected for high school)</i>	<i>Chemistry Median (Corrected for h. s. & intelligence)</i>
A	M.S.	Instructor	2	14.5	110.4	84.0	75.5	72.27
B	Ph.D.	Asst. Prof.	5	14	105.6	81.5	76.25	74.55
C	B.S.	Instructor	2	10	127.3	105	88	79.36
D	M.A.	Instructor	1	12.5	118.8	77.5	63.83	57.91
E	M.S.	Instructor	5	21	107.7	65	57	54.63
F	B.S.	Instructor	3	20	113.2	79	68	63.87
G	B.A.	Instructor	4	18	107.6	84.5	76.5	74.16
H	B.S.	Instructor	15	21	104.9	75	64	62.53
I	M.S.	Instructor	3	21	117.4	83.5	77.25	71.78
J	Ph.D.	Assoc. Prof.	14	15	101.3	66.75	59.5	59.18
K	B.S.	Instructor	7	18	80	69	68	74.5
L	M.A.	Instructor	2	21	115.7	91	80	75.07
Averages					109.6	78.04	71.07	68.09

^a Moss, F. A., Loman, Wm., and Hunt, Thelma, "Impersonal Measurement of Teaching," *The Educational Record*, Vol. 10, No. 1, January 1929, p. 47.

structors' classes in intelligence, he would have .32 added to his chemistry average. The data obtained for twelve instructors all teaching general chemistry classes in one college are shown in Table XXIII. Similar records were available for all the instructors studied. The last column of the table contains the final figures representing relative efficiency of the instructors. By reference to this column, it can be seen that the variations in efficiency of instruction even within the same institution are large. Among all the instructors in all the colleges included in the study, the efficiencies as indicated by corrected median scores on the chemistry test range from 37 to 79.

It may be said that other factors not taken into account in this study affect the teaching efficiency, and so lower the reliability and validity of estimates based upon student test performances. Studies done in conjunction with the investigation just described showed relatively little effect, however, from such factors as departmental organization, size of institution, size of class, and teaching load of the instructor. While this method of measuring teaching efficiency may not be as easily applicable to all types of classes as it is to chemistry, it certainly suggests principles along which efforts may be made to measure efficiency of instruction in those courses of study for which the content is fairly well standardized. It is a method of demonstrating results of teaching objectively; and one which would remove rating of teacher efficiency from the realm of subjective personal judgment.

PART V

**MEASUREMENT IN INDUSTRIAL AND
PERSONNEL FIELDS**

CHAPTER XV

Historical Background of Psychological Measurement in Industry

PSYCHOLOGICAL measurements in business and industry are those measurements which are applied to the human element and to the processes directly dependent upon the human element. In the early days of large-scale business and industry, psychological measurements found relatively little application, because of the lack of emphasis upon man as a part of the production process. The early industrialists thought of the machine as the all-important element in business. The introduction of psychology and psychological measurements into industry represents an admission that the machine is insufficient in itself to meet the demands of industry. The study and measurement of the human element marks a recognition that machines can be used to advantage only through properly selected and adequately trained men.

I. The Foundations of Psychology Applied to Personnel Problems

In his discussion of the historical background of industrial psychology, Viteles points out that three distinct forces have played a part. These he refers to as the *economic*, *social* and *psychological* foundations of industrial psychology. Since industrial psychology is largely the application of quantitative methods to the study of

258 History of Measurement in Industry

human traits and abilities and working processes, we may consider these three forces as the fundamental ones leading to the application of psychological measurement to the problems of personnel. The *economic foundation* is the aim of business and industry to attain maximum production at minimum cost. It was to be expected that whatever might contribute to the attainment of this aim would be accepted by industry. As a matter of fact, industry has accepted psychology and psychological measurement as aids largely because of what they can contribute toward cheaper production and greater profits. One of the most important reasons why industry is willing to introduce measurements of ability, of aptitudes, and of personality as a means of selecting employees, is that employees selected on such a basis are worth more to them than employees selected on other bases.

Antedating somewhat the specific applications of psychology among workers were a number of early studies which directed attention toward the human element in industrial management and demonstrated that such consideration was economical. These included the studies in scientific management conducted by such workers as Frederick W. Taylor and Frank B. Gilbreth. Taylor, toward the end of the nineteenth century, started a system of scientific management which was the first to emphasize the human factor in production. He became the pioneer of a movement which spread throughout the entire world of industry under the name of Taylorism. His system was based upon two assumptions: ¹ (1) what the workmen want from their employers beyond anything else is high wages, and what the employers want from their workmen is a low labor cost of manufacture;

¹ Taylor, F. W., *The Principles of Scientific Management*, New York, 1911, p. 42.

and (2) no system or scheme of management should be considered which does not in the long run give satisfaction to both employer and employee, which does not make it apparent that their best interests are mutual, and which does not bring about such thorough and hearty coöperation that they can pull together instead of apart. Taylor made an early application of his principles in his classic experiment in selecting and training men to handle pig iron. This experiment has been well summarized by Viteles: ²

The work was carried on in the plant of the Bethlehem Steel Company. At the time of the experiment there were in operation five blast furnaces, the product of which had been handled by a pig-iron gang for many years. This gang, at the time Taylor started his work, consisted of about 75 men. "They were good, average pig-iron handlers, were under an excellent foreman who himself had been a pig-iron handler, and the work was done, on the whole, about as fast and as cheaply as it was anywhere else at that time." The work of handling pig-iron was done by men with no other implements than their hands. The pig-iron handler stooped down, picked up a pig weighing about 92 pounds, walked for a few feet or yards, and then dropped it onto the ground, or upon a pile. Taylor believed this work to be so crude and elementary in its nature that it would be possible to train an intelligent gorilla to become a more efficient pig-iron handler than any man can be. And yet he felt that the science of handling pig-iron, in spite of the crude character of the work, is so great, that this type of work could be used to illustrate the accomplishment to be derived from properly training competent workers in the best methods of work.

When Taylor started to apply his principles, he found

² Viteles, Morris S., *Industrial Psychology*, W. W. Norton & Co., Inc., New York, p. 10.

that the gang of laborers employed in loading pig-iron onto a railroad car were averaging approximately $12\frac{1}{2}$ long tons per man per day. After carefully observing the methods of work and studying the number of voluntary pauses, etc., he reached the conclusion that a first class pig-iron handler ought to handle between 47 and 48 tons per day instead of $12\frac{1}{2}$ tons. Taylor questioned many good managers and asked them whether, under premium work, piece work, or any of the ordinary plans of management, they would be likely ever to approximate 47 tons per day. Not a man suggested that an output of over 18 to 25 tons could be obtained by any of the ordinary expedients.

Taylor then set about to choose a worker on whom he could first try out his new methods. He finally selected a Pennsylvania Dutchman whose reputation, habits, and ambition made him seem a likely subject. This man was asked whether he would prefer earning \$1.85 per day to the \$1.15 which constituted his pay check at the time. He was told he could do so by loading in one day a pile of pig-iron (consisting of $47\frac{1}{2}$ tons) which was pointed out to him. He was further cautioned that in order to load this pile of pig-iron and to earn his increased pay, he must carefully follow the instructions of the man assigned to train him in the proper method of doing his work. He started to work, and all day long, and at regular intervals, he was told by the man who stood over him with a watch; "Now pick up a pig and walk. Now sit down and rest. Now walk—now rest," etc. He worked when he was told to work, and rested when he was told to rest, and at half past five in the afternoon had his $47\frac{1}{2}$ tons loaded on the car. He practically never failed to work at this pace and to do the task that was set him during the three years of observation by Taylor. And throughout this time he averaged a little more than \$1.85 per day, whereas before he had never received over \$1.15 per day, which was the ruling rate of wages at that time in Bethlehem.

One man after another was picked out and trained to handle pig-iron at the rate of $47\frac{1}{2}$ tons per day until all

the pig-iron was handled at this rate, and all of the men still working in the gang received 60% more wages than those paid to other men around them who were not employed on task work. However, of the gang of "seventy-five pig-iron handlers," only about one man in eight was found physically capable of handling $47\frac{1}{2}$ tons per day. With the very best of intentions, the other seven out of eight men were physically unable to work at this pace.

Such experiments as the one just quoted are excellent examples of the demonstrations to industry that attention to the human element is economically profitable. While Taylor and other industrial engineers who were interested in his principles contributed very little to the theory and procedure of psychology as such, their emphasis on the human factor was extremely important in preparing the ground for later application of psychology.

The *social foundations* of industrial psychology are rooted in development of an attitude of concern for the well-being of the worker. At the beginning of the industrial era, the worker was looked upon largely as a commodity with a certain commodity value, little thought being given to his individual happiness and well-being. During the nineteenth century, when development in industry was proceeding so rapidly, the necessity of protecting and conserving the human element was rarely recognized. Attention was given to the conservation of capital, machinery, and raw materials, but waste of human life, the devastating effects of fatigue, and destruction of health in industry received very little attention. Recent years have brought a considerable change in this viewpoint and today individual welfare in industry is acknowledged to be an important concern of society. The concern is reflected in all sorts of social welfare measures, compensation for accidents, pension

systems, health protection, etc. This attention to the individual welfare of the worker has finally culminated in the development in large organizations of separate personnel departments, usually manned with individuals trained in psychology and equipped with the various instruments for studying and measuring aspects of human abilities and skills.

The *psychological foundations* of industrial and personnel psychology are rooted in the development of psychological techniques of aid in dealing with the worker or employee. The introduction of experimental and quantitative methods into psychology marks the beginning of the development of techniques which can be utilized in solving the problems of the worker. The study of individual differences has also been fundamental to the development of industrial psychology. In fact, industrial psychology is concerned with the individual—in his reactions in a specific vocational situation. More recently, the development of the various types of psychological tests has contributed immensely to the solution of problems of employment.

II. Early Experiments in Industrial Psychology

The earliest applications of quantitative methods of studying the worker were scattered experiments done in psychological laboratories. Some of the early investigators became interested in problems of fatigue among workers and conducted various types of studies in which attempts were made to measure this fatigue. Others of the early investigators were interested in studies related to the training of workers. Of significance are the investigations of Bryan and Harter on the learning of telegraphy, and of Book on learning to typewrite. As early as 1911, Walter Dill Scott published a monograph

in which he discussed the applications of psychological devices in recruiting employees and in increasing the quantity and quality of their work.

While many of these early experiments had a relationship to the application of psychology in industry, no systematic attempt to apply *psychological tests or measurements* to a practical industrial problem was undertaken before the pioneer work of Münsterberg in 1912. At that time he undertook an experiment in scientific selection of street-car motormen at the suggestion of and with the coöperation of the American Association for Labor Legislation. Those interested in the experiment were concerned mainly with the problems of safety and accident prevention. They recognized the importance of the human element in the causation of accidents and were therefore willing to sanction a psychological study of the problem. Münsterberg undertook to analyze the human qualities important in accident prevention and to devise tests for measuring the qualities. After careful study, Münsterberg stated that one of the most necessary qualities in the safe operation of the street car was "a particular complicated act of attention by which the manifoldness of objects, the pedestrians, the carriages, and the automobiles are observed with reference to rapidity and direction in the quickly changing panorama of the street." He emphasized as of extreme importance "the ability to keep attention constant, to resist distraction by chance happenings on the street, and especially the always needed ability to foresee the possible movements of pedestrians and vehicles."

For these qualities, Münsterberg devised a test consisting of a series of cards representing a street with the various conditions which might exist with reference to pedestrians and vehicles. The arrangement of a sample

264 History of Measurement in Industry

card is shown in Fig. 19. The heavy lines represent street car tracks. The space on either side of the track is divided into 64 units, represented by small squares. The "1" digits represent pedestrians, who move just one step, from one unit to the next. The "2" digits represent

			2	A				
			3	B	1			
	3	1	2	C	1		2	
		2		D	2		3	
	2		1	E		1	1	
	3		1	F	1	3		2
		2		G	2	3	2	
	3		3	H		2		3

Fig. 19.—Sample Card from Munsterberg's Test for Street-Car Motormen. Large, heavy figures represent red digits.

horses, which move twice as fast, or two units. The "3" digits are automobiles, moving three units. Black digits move parallel to the track and are, therefore, not sources of danger in causing street car accidents. Red digits stand for pedestrians or vehicles which move toward the track and are potential sources of danger. Those red digits

which would land in the track in taking the number of steps indicated are considered dangerous ones, and in taking the test the subject must pick out as quickly as possible those points on the tracks that are so threatened. As the test is taken, the cards are exposed by a moving belt with a window that exposes the whole width of the card and an area of 5 units' length. The subject turns a crank moving the belt as he takes the test. The score on the test is determined by errors and time taken to observe. Münsterberg gave the test to a group of motor-men of a street railway company and found considerable relationship between scores made and accident records.

This pioneer experiment of Münsterberg was the starting point for a number of trials of his test, for the development of a number of similar tests, and, in its far-reaching effects, for the beginning of a widespread coöperation between the science of psychology and industry in the solution of industrial problems.

III. The Effect of the World War

The development of psychological measurements for recruits in the United States Army during the World War represented the largest-scale experiment in the application of psychological methods to an industrial problem that we have ever had in this country. The work of developing the mental tests is described elsewhere in this book.³ In addition to the well-known mental tests, investigations were conducted in the development of specialized tests for aviation and for other more technical occupations. There were developed not only measures of general capacity, but a series of trade tests for measuring the skill of men being considered for assignment to various special

³ See Chapters III and IV.

skilled trades, such as electrical work. These various developments in the Army were the starting point for many experiments on scientific selection for various occupations which were carried on after the war in civilian life. In fact, the war testing probably gave the impetus to the rapid rise of selection methods, for workers, based upon the testing of capacities and skills.

IV. History of the Testing of Applicants

Since the greatest, or at least the most extensively used, contribution of psychology to personnel problems has been in the field of employment examinations, our historical sketch might well include a view of the testing of applicants. Testing by examinations on an extensive scale began in the public service in connection with the selection of government employees. The setting up of examinations in the public service dates generally from the establishment of the United States Civil Service Commission. This Commission was established by law in 1883 to select employees on a basis of merit, to supplant the old spoils system under which employees held their jobs largely through political influences.

Since the founding of the Civil Service Commission there has been a continuous evolution of the system of testing applicants, starting with the borrowing of purely academic examinations from the schools, and culminating in the extensive work on psychological tests carried on by the Research Division of the Civil Service Commission. The stages of the evolution may be represented by the following types of tests: (1) examinations in terms of school subjects; (2) essays in terms of the job; (3) general intelligence tests; (4) special aptitude tests in terms of the job; (5) achievement or trade tests in objective or short-answer form.

1. **Examinations in terms of school subjects.** When the Civil Service Commission first faced the problem of selecting employees on the basis of merit, practically nothing had been done in the development of examination methods which might indicate degrees of merit. The only types of examinations which the Civil Service Commissioners found available were the academic examinations used by educators as a basis for judging performance and merit in school work. No other examining method being available, those given the job of selecting government personnel under the new Civil Service Act began using these academic tests for all kinds of positions. The tests usually consisted of parts covering such subjects as arithmetic, spelling, composition ability, penmanship, and grammar. The tests varied from one job to another only in the difficulty of the questions. The same types of academic tests were used to fill such diverse positions as clerk, librarian, matron, patrolman, and messenger boy.

As we view this type of examination today, the only merit which it might possess as a basis for selecting such diversified employees would be in its selective power with reference to general ability or general mentality. Presumably, those who used the examinations might have believed that high test attainments were indicative of high general ability. However, numerous experimental studies have demonstrated that there is relatively low relationship between one's penmanship ability and one's mental alertness, or between one's spelling ability and one's mental alertness. About the only thing that we can say in favor of the purely academic tests is that they were undoubtedly better than selections based purely upon personal opinion gained from an interview or some other similar method.

2. Essays in terms of the job. This stage in the evolution of methods of examining applicants grew immediately out of the objection to the academic tests on the score that they were unrelated to the jobs to be filled. In the first attempts to overcome this objection, we find Civil Service examiners using the same examination techniques but asking questions having some bearing on the job to be filled. If a composition was to be written, the subjects suggested to the applicant related specifically to the job, instead of being of a general nature. In this stage of development, the examinations often included a specific part on the job; if the job was that of plumber, some of the questions asked related specifically to plumbing. But since this stage antedated the development of objective, short-answer types of tests, the questions were of a discussion or essay type.

The chief advantage of this improvement in examination methods lies in the more specific bearing of the questions on the job, which made them appeal to competitors and employers as being practical. Disadvantages were chiefly those which we pointed out in Chapter XII as existing in the case of essay tests as compared with more objective test methods. This method of selecting employees combined with the purely academic tests continued for several years and still exists in many instances. Practically no other types of tests appeared before the development of psychological test methods during the World War.

3. General intelligence tests. The advisability of using general intelligence tests for determining mental alertness having been demonstrated by the experience in the Army during the World War, various civil service commissions and private industrial groups quickly saw the possibilities of utilizing such tests in the selection of

employees. At the beginning most of the testing was done by using the actual tests which had been developed in the Army. Later on, with the realization that tests could be developed better adapted to the specific problems at hand, general intelligence tests designed particularly for occupational use were more frequently worked out. There are scattered instances of such developments in private industrial groups. With the establishment of the Research Division of the United States Civil Service Commission in 1922, there began the development of a series of mental alertness tests for use in selecting public personnel. For a discussion of the advantages and limitations of general intelligence testing in the selection of employees, the reader is referred to Chapter VII.

4. **Special aptitude tests in terms of the job.** The use of general intelligence tests was a distinct advance over the old academic types of tests for selecting personnel, but such tests possessed certain disadvantages which led in a short time to the development of specialized types of tests. The intelligence tests being general in nature, their terminology often seemed rather "far-fetched" in relation to the occupation for which they were used, even though they did test a quality important in the job. This made them in many instances seem abstract and impractical to both applicant and employer. It also proved to be true that for a considerable number of jobs the relationship between abstract intelligence and success on the job was not high enough to justify putting complete dependence on the general test for the selection of employees. Such facts as these led to the construction of tests patterned largely after the general intelligence tests but more specifically applying to the position for which they were to be used. In some instances these special tests are nothing more than specialized intelli-

270 History of Measurement in Industry

gence tests in the terminology of the job. In other instances, they measure certain additional qualities not measured in the general mental alertness test. Specialized tests have the advantages of appealing to administrators as being practical, of appealing to the applicant as being a fair test of the ability required of him, and of having generally a better selective value or validity than the abstract intelligence tests.

An example of one of these specialized intelligence tests is one constructed by Telford and Moss for selecting policemen. This particular test has been used extensively by state and city Civil Service Commissions. The whole examination includes the following parts:

(1) *Observation*. Measured by presenting the applicant with a picture of a collision between a street car and an automobile, requiring him to study it for a limited time, and asking him later to answer several questions without looking at the picture.

(2) *Memory*. Measured by requiring the applicant to pick from a large number of photographs faces that he has seen before.

(3) *Comprehension*. Measured by the applicant's ability to answer questions based on printed selections from laws, ordinances, and police regulations.

(4) *Judgment*. Measured by ability to answer such questions as the following,⁴ in which the applicant has to select the correct one of four suggested solutions:

If a policeman considers himself unfairly treated by his Sergeant, and gets no satisfaction when he explains to the Sergeant that he is not treated fairly, he should:

-----Refuse to obey any orders given by the Sergeant.

⁴ Telford, Fred, and Moss, F. A., "Suggested Tests for Patrolmen," *Public Personnel Studies*, 1924, Vol. II, p. 112.

- Invite the Sergeant to meet him when both are off duty so that they can settle the matter themselves.
- At the first opportunity report the matter to his Lieutenant or Captain.
- Immediately hand in his resignation.

It will be noted that this test as compared with a general intelligence test stresses those things which the applicant might be expected to do after he assumes his job. His observation is tested in terms of a type of situation which he may actually be called upon to observe and report; his memory is tested in terms of faces, a type of memory which the policeman must often utilize in the recognition of persons to be apprehended; and his comprehension ability is tested in terms of material related to the job.

5. Achievement or trade tests in objective form. These tests have been developed for positions in which high intelligence or special aptitude alone is not sufficient, but in which the applicant must have specific information and technical training along the line of the job itself. The achievement tests are designed to measure acquired information or knowledge about the job or technical skill. They are an outgrowth of the essay tests on the job, coming with the demonstration that objective methods have many advantages over essay methods in the measurement of achievement.

Short-answer tests which have been developed for use in selecting chemists, bacteriologists, hospital workers, etc., are illustrations of this type of examination. The trade tests developed during the World War for measuring knowledge of mechanical and electrical work are also examples. These tests are primarily achievement tests and, in industry and personnel work, serve the purposes

272 History of Measurement in Industry

that educational achievement tests serve in the schools. Unlike general intelligence tests and aptitude tests, they are designed not simply to indicate ability to learn a job, but to measure the amount the applicant actually knows about the job at the time the test is taken.

CHAPTER XVI

Constructing a Psychological Test for Employment

BEFORE discussing specifically any of the tests which have been devised for use in industry, we shall consider the steps involved in the construction of a psychological test for use in this field. This is a very important part of a consideration of industrial tests, since the difficulties of carrying out these steps have constituted the chief drawbacks to a wider application of psychological measurements in dealing with problems of employee selection and control. Many of the points to be outlined have already been mentioned in the discussions of tests in the other fields. The importance of test validity, for example, and the importance of test reliability have been emphasized many times. These are two criteria for good tests which must be met no matter what the field of testing is. It is also true that many of the other procedures to be discussed in this chapter are just as applicable to tests in the field of school achievement, or in the field of intelligence testing, or in the field of personality testing. Because they are somewhat less likely to be understood and somewhat more likely to be neglected by the industrialist, we are emphasizing them here.

Since most of the psychological tests of industry have been designed primarily for selecting employees or for

indicating abilities which make for success on certain jobs, our outline of the steps in constructing an industrial test will be presented with such a test in mind. Also, since the contribution of psychologists to measurement in the selection of employees has most often been in the form of pencil-and-paper tests, similar in their general plan and structure to the pencil-and-paper intelligence tests so frequently applied in the academic world, our outline will fit best such a test. It should, however, be kept in mind that the general principles apply whether the test be for selection or for control of employees, or whether the test be of a pencil-and-paper type or of some manual-performance type.

The steps necessary in the construction of an industrial test may be briefly outlined as follows:

1. An analysis of the duties of the job and the qualifications necessary for performing the job.
2. The selection of types of tests and types of questions to be used.
3. Decision on approximate length of the test.
4. The construction of the test questions.
5. The administration of the questions to a trial group, and the analysis of the test material on the basis of the trial.
6. The selection and arrangement of final parts and questions to be used.
7. The application of the test to groups for whom the test is designed.
8. The establishment of norms and critical scores.

These steps as outlined carry the problem all the way from the original need for a test to the final establishment of critical scores for making use of the test. A step which might be added to the above is one involving a survey

and study of already available material related to the problem. In any test construction job, this should precede the development or study of new material. What has already been done on related tests may give valuable information on a new problem. We shall now consider in somewhat more detail the various steps outlined above.

1. **Analysis of the job.** It is quite obvious that it would be impossible to construct a measuring device for selecting people for a job without knowing anything about the job. The personnel psychologist, therefore, sets out to acquire a knowledge of the duties of the job and a knowledge of the qualifications necessary to perform the duties, the latter probably being the more important to know. As he approaches his problem, he may find that analyses have already been made of the job. This is true in a great many of our state and city civil-service organizations. Classification boards have in many instances undertaken to analyze all of the jobs coming under the state or city jurisdiction, and these analyses may be summarized for each job studied. Or, the psychologist working in a private industry, particularly if it be a large organization, may find a classification plan with an outline of the duties and qualifications for the various jobs involved.

On the other hand, the psychologist who constructs the measuring device may be faced with the problem of making his own job analysis. He undertakes "a process of dissecting a job and describing its component elements." The jobs which must be performed and the steps necessary for the performance of each of the operations are described in detail. The qualifications for the job are also studied, and such things as the working conditions, the incentives, and the morale may be included. The

analysis of qualifications often includes an enumeration of the various mental traits which are necessary for the job.

These analyses may be made by the tester in several different ways. Probably the best way is by a direct study of employees actually on the job; that is, a study involving actual observation of working conditions and working procedures. In other instances, job analyses may be based upon reports of executives and foremen, their opinions being gained in interviews with them or through questionnaires which they may answer. Questionnaire methods of analyzing jobs have also been utilized with the employees themselves.

To summarize the chief purposes of the job analysis, so far as the constructing of psychological measuring devices is concerned, we may say that the analysis (1) gives a basis for selecting the types of tests to be used, and (2) indicates the general level of ability to be aimed at in tests.

2. Selection of types of tests and questions. The selection of types of tests may be viewed from two standpoints. We may think of the selection of tests in terms of qualities which they measure, in which case we may consider memory tests, reasoning tests, judgment tests, information tests, tests of observation, or tests of speed. On the other hand, we may think of types of tests in terms of whether they are to consist of true-false statements, multiple-choice questions, questions to be answered by single words, columns of items to be matched, etc. In most instances, the first step is a selection of type of test from the standpoint of quality to be measured, this selection being based upon job analyses or job specifications. The type of test from the standpoint of nature of the questions is usually a subsequent pro-

cedure secondary to the qualities to be measured, the types of questions being selected on the basis of their suitability for measuring the qualities or traits selected for measurement.

3. **Length of test.** It is important to make a preliminary decision as to the length of the test at an early stage in the procedure, since this will determine the number of questions to be constructed, the obtaining of trial groups, and a great many details of the further study of the material. The decision as to length should depend primarily upon reliability and validity. The length should be sufficient to insure a reliable test of the qualities to be measured, and yet should not be so long as to interfere with the validity through factors of fatigue or boredom. In the practical situation, it often happens that factors of available time for giving tests, available help for scoring tests, and available money for administering the tests, have to be considered. In an ideal situation, however, these should be secondary to the factors that make a good test.

4. **Construction of questions.** This step involves the actual making of the question material. The number of questions to be constructed is determined by the desired length of the test; the general nature of the questions to be constructed is determined by the qualities to be measured and the level of ability of those to be tested. The subject-matter content of the questions is determined by the specific job for which the test is being constructed. Ordinarily, a number of questions somewhat in excess of the final number desired is constructed. This is to allow for elimination of the questions which prove to be poor testing material in the preliminary trials. So far as the nature and subject matter of the questions is concerned, in the initial construction we can

be guided only by our job analysis and by previous experience with questions in the same or a similar field.

5. Analysis of test material from preliminary trial. After the material for a given test has been constructed, it should be administered to a trial group composed of individuals of known ability on the job. This usually involves giving the test to a trial group of individuals already on the job. Such a trial gives a basis for studying the various parts of the test and the various separate questions in these parts with a view to determining their actual value in differentiating the good and poor or the efficient and inefficient employees. It is customary to study the parts of the total examination for (1) *validity*—relationship to efficiency of the employees of the trial group; (2) *interrelationships* among the various parts of the test; and (3) *distribution of scores* or grades given by the various parts. Ideally, each part of the test shows a high relationship between scores made on it and job efficiency; shows a good distribution of scores (not all bunched within a few points of each other); and is not so closely related to other parts as to be measuring the same thing. The separate questions of the tests are usually studied from the standpoint of difficulty and “selective value” or question validity. Each question should be studied for its difficulty by a tabulation of the number of people in the trial group who answer it correctly or incorrectly; and each question should be further studied for validity or selective value by a tabulation of correct and incorrect answers made by a group of good employees as contrasted with a group of poor employees. The last of these analyses, that for selective value, can usually be made by dividing the trial group into halves, thirds, or quarters on the basis of their efficiency on the job. After such an analysis, questions are retained for the test

which show a difficulty within suitable range for the group to be tested, and which show a small proportion of errors for the good group and a large proportion for the poor group.

6. Final arrangement of test. On the basis of the analyses just discussed, the final test is put together. Those parts are selected which show the highest validity. Those questions are selected which prove to be of suitable difficulty and suitable selective value. At this point, also, attention is given to the sequence of the various parts of the total examination, as well as to the sequence of individual questions within a part. The sequence of parts of a test is determined largely on the basis of time required for answering, difficulty, and appeal to the person being tested. The sequence of the individual questions within a part is usually a difficulty sequence, questions being arranged in order of increasing difficulty.

7. Final application of test and establishment of critical scores. Before a test can be applied in the actual selection of employees, we should know the significance of various scores on the test. In other words, we should be able to answer the question, "At what score level should a person be accepted for employment or rejected for employment?" This involves the application of the test in its final form to groups of known ability, and from such an application, the experimental determination of the points below which employees are unsuccessful. Such determinations are ordinarily spoken of, particularly with reference to academic tests, as the establishment of "norms" on the test. With reference to the industrial test, we more frequently speak of it as the establishing of critical scores, or "passing" points.

CHAPTER XVII

Measurement in the Selection of Employees

I. The Purpose of Measurement in Selection

MEASUREMENT in the selection of employees is generally aimed at selecting before employment those who will be efficient on the job, who will be relatively satisfied and happy on the job, and who will remain with the job long enough to make it worth while to train them into the work. Stated in another way, measurement in the selection of employees has for its main purpose the obtaining of workers who will efficiently and economically carry out the work of the organization of which they are a part. Just as the mechanical engineer wishes to know about the qualities of a machine before purchasing it, it is logical to expect that human engineers—personnel directors, executives, and industrial managers—will want to know about the qualities of an applicant before employing him, at least about those qualities which will determine his success. We have already considered at another point in this text the qualities in which the human engineer is most interested.¹ These are, briefly, the natural abilities and special capacities requisite for performance on the job, the acquired skills and achievements necessary to success, and in some jobs particular

¹ See Chapter VII.

personality traits. The primary purpose of any selection procedure is to differentiate between those who do and those who do not possess these qualifications. In addition to eliminating those who do not meet the minimum requirements of the job, it is desirable that selection methods grade those who possess the qualifications in varying degrees.

Both experience and experimental analysis of traditional procedures employed in the selection of workers have shown that in many cases these procedures fail to accomplish their purposes. This has made employment managers and personnel directors alert to any improvements in techniques of measuring human qualities which will be of advantage in differentiating the good from the poor employee. All of the psychological measuring devices described in this text are of interest in relation to problems of selection. Almost all of them have at one time or another been tried, at least in an experimental way, in relation to the selection of employees, and many of them have been actually introduced as a permanent part of selection procedures. The lower susceptibility to error of tests of ability and achievement has made these tests the most used among the various psychological measurements in the selection of employees.

II. Problems Related to Measurement in the Selection of Employees

All of the problems which are generally involved in the construction and evaluation of psychological measuring instruments are, of course, involved in relation to their development for use in employee selection. Certain of the problems, however, are of particular importance in the application of measurement to personnel work and, therefore, deserve special mention at this point.

1. **Analysis of the job.** As was pointed out in Chapter XVI, the first step in fitting men to jobs and in developing measurement procedures for selecting the men is the making of a comprehensive study of job activities and requirements. Such an analysis is basic to the working out of any test or any other measuring instrument for predicting vocational success. It cannot be doubted that some selection programs have failed because they have not been founded upon scientific and adequate job analyses. The shortcomings of job analyses which are to be used as a basis of developing employment tests and measurements are most likely to rest upon the following:

(1) *The analyses are often too hastily made and too superficial in nature.* (2) *Poor methods may be employed in obtaining the information to be used as a basis for the analyses.* Too often the information is gathered from foremen or supervisors who do not have an intimate enough contact with the actual job to be able to supply all the information necessary. In other instances, deficiencies may be due to the utilization of poorly constructed questionnaires for obtaining information from employees on the job. (3) *Persons making job analyses may not be qualified to undertake such procedures.* Best results would seem to demand that analyses be made by persons trained in personnel work, persons with an adequate understanding of human traits and abilities, and persons at least somewhat familiar with the job processes to be studied. (4) *A final shortcoming in job analyses is often the failure to arrive at traits and qualifications necessary for success.* Too many analyses are limited almost exclusively to statements of the duties and the working conditions with no indication of the mental capacities and other traits essential for success on the job.

It is to be recognized, of course, that the analysis of the capacities and traits essential for success is a difficult task, but is much more essential to the development of measurements for selecting employees than are mere statements of duties and working conditions.

These various difficulties which arise in the making of job analyses are mentioned because of the fundamental nature of these analyses in the development of good tests and measurements for the selection of new employees. If it is not based upon definite knowledge of the job for which it is to be used, it is only by chance that a selecting instrument turns out to be useful.

2. Problems of test construction. The general problems of test construction were covered in Chapter XVI; they will not be repeated here. There have been some instances of difficulties arising on this score in the development of employment tests because persons untrained in the technique of psychological measurement have attempted to construct the tests. In most of these instances, the tests have been doomed to failure, and they have often had a more serious and far-reaching effect in that they have discouraged the industrial organization from general extended use of objective test procedures. It is encouraging to find that there is today, on the part of industrial organizations and personnel agencies in the public service, a growing demand for individuals trained in the construction and evaluation of objective test procedures.

A second problem in relation to construction of employment tests is that of inability to construct instruments of measurement which will reliably measure all of the traits desirable in an employee. The field of personality measurement, for example, has not yet reached a

point at which it can generally supply reliable and practical instruments for measuring personality traits in prospective employees.

3. **Establishing criteria for evaluating measurements in selection.** The evaluation of employment tests and other selection methods involves a comparison between the test scores and the success of workers on the job. There is not a real basis for the introduction of a new selection procedure unless it can be demonstrated that the procedure actually measures success on the job, and the only way to do this is to have available a standard of accomplishment on the job. This standard of accomplishment on the job is usually known as the *criterion of vocational success*, so-called because it is the criterion by which we judge the efficiency of our test or other measuring instrument. Many studies of tests have failed or have arrived at no definite conclusions because an adequate criterion of success was not available or was not established. To quote from a discussion of this point by Bingham,

many a study of methods of selecting people for jobs has led to ambiguous conclusions because of the inadequacy or unreliability of the criterion by which the methods were judged. All too often a research has passed through the laborious and expensive phases of making the job analysis, constructing ingenious tests, and giving the tests to numerous employees, before the investigator discovered that no adequate and reliable measure of individual achievement on the job was to be had.²

Criteria of vocational success are of various types. Viteles has divided them into two groups: *objective* and

² Bingham, W. V., "Measures of Occupational Success," *Harvard Business Review*, 5 (1926), pp. 1-10.

subjective. The former includes factors which can be definitely measured and expressed in objective terms. They may be quantity of output, quality of output, number and cost of accidents or breakage, length of service on the job, rate of advancement in the job, compensation where it has a definite relationship to production, and performances on standard trade or job achievement examinations. Objective criteria are to be desired in all programs of evaluating selection procedures, but in many instances they are impossible of attainment. Too often quantity and quality of output are not measurable in any simple quantitative terms. We might think, for example, of the relatively simple job of the clerical worker. His work is practically always of too varied a nature to be measured directly in terms of amount done. As a rule, only for those jobs in which the individual is working on a definite piece-rate basis will it be possible to establish an adequate criterion in terms of direct measures of quantity and quality of output. The *subjective criteria* are usually criteria based upon supervisors' ratings or estimations of success and efficiency. Too often these subjective criteria are so unreliable that comparisons of test procedures with them mean little, if anything.

4. **Convincing employers and managers of the value of measurement in selection.** In the introduction of psychological measurements into employee selection, psychologists have had to do a great deal of pioneer work in convincing industrial managers, executives, and others in charge of employee management. Until recently, very few executives and industrial managers have been trained in the test techniques which we have been considering, and they have been slow to adopt the measuring instruments which have more frequently been worked out in

academic circles. As we pointed out in the historical consideration of psychological measurement in industry, the new measuring devices usually have had to await a demonstration that they would be of economic advantage to the employer.

III. Measurement in the Selection of Employees Illustrated

We noticed in our discussion of the use of mental tests (Chapter VII) that the public service has generally been somewhat in advance of private business and industry in the utilization of the newer methods of employee measurement. The United States Civil Service Commission, through its Research Division and through the efforts of the director of this division (Dr. L. J. O'Rourke), has been outstanding in accomplishment in this respect. We shall examine some of its work as an illustration of scientific test development in selecting employees. The background of its work is of sufficient interest to warrant our noting it. The following is from a report by Herbert A. Filer, at one time Chief Examiner for the Commission:

Notwithstanding the financial handicaps under which the Commission has always found itself, it has constantly endeavored to reach out and make use of improved methods wherever they could be found. In January, 1917, learning of work being done by Dr. Edward L. Thorndike of Columbia University and Dr. Walter Dill Scott of Northwestern University, in devising tests for employment purposes, Mr. George R. Wales, then chief examiner of the Commission, wrote to these two men asking them for permission to see a complete set of the tests they had developed and requesting any comment that they might deem of interest

in connection with the Commission's work. Dr. Thorndike replied that Dr. L. K. Frankel of the Metropolitan Life Insurance Company could furnish copies of the tests arranged for clerical workers; and upon request Dr. Frankel did so, with the understanding that the copies of the tests were furnished for confidential sight only and were to be returned to him after inspection. Dr. Scott replied, enclosing a preliminary announcement and two reprints of tests, stating that it would be seen that the work was simply experimental, and that it was desired that none of the material should be given out until it had been thoroughly tested.

It will be recalled that war with Germany was formally declared on April 6, 1917, less than three months after this correspondence was initiated. Before and after the declaration of war, the Commission was literally swamped with work incident to recruiting the civilian forces of the Government; hence no opportunity was afforded to accept an invitation extended by Dr. Scott to visit him and inspect his work; there was no time to follow up the subject in any direction.

Soon after the armistice the Commission arranged with Dr. R. M. Yerkes to give the Army Alpha test to about one hundred of its own clerical employees. The results of this test were studied and the relationship found between the Alpha scores and the individual efficiency ratings of the employees, as well as between the entrance examinations of these employees and their efficiency. While this trial did not show that the Army Alpha was an improvement over the Commission's established examinations for selecting clerical workers, it did indicate that tests might be devised according to the basic principles of the Alpha and similar tests, which would be advantageous for Government employment purposes.

In April, 1919, Dr. John B. Watson, head of the Department of Psychology of Johns Hopkins Hospital, was engaged by the Commission as an expert examiner to start experiments with a view to determining whether the principles observed by psychologists in testing work

were applicable to civil-service examinations, and if so, to what extent. The experiments were made in the Baltimore post office, tests being given to several large groups of clerks employed in that office. The results indicated that neither the experimental tests nor those given by the Commission for entrance, bore a sufficiently close relation to the efficiency ratings as furnished by the post office authorities to establish definitely their selective qualities. Either the tests were not appropriate or the efficiency ratings were incorrect, and it seemed probable that both were faulty.

Dr. Watson then frankly stated that he would not have time to devote to continued research for the Commission, but that he was interested in the "merit system" of employment and recommended that a psychologist engaged in such work be consulted with a view to employment by the Commission as consulting examiner. He mentioned the name of Dr. Beardsley Ruml, then associated with Dr. Walter Dill Scott in the Scott company. Accordingly, Dr. Ruml was communicated with and came to Washington for the purpose of making a survey of the Commission's examination methods, with a view to formulating a constructive program for the improvement of those methods.

Dr. Ruml's report was dated June 10, 1920. He recommended in substance that, in order to make a saving in the current expenditures of the Commission sufficient to provide for the establishment of a small research unit, a rearrangement be made of some of the tests used in the more generally attended examinations, such as those for the Postal Service, so as to economize in the scoring and other handling of the papers without making radical changes in the content of the examinations themselves. He pointed out that upon the establishment of a research unit the efficiency of examinations in use could be checked against the demonstrated proficiency of employees, and that new and different methods could be substituted wherever their superiority could be proved. This program seemed wise to the Commission, and it was adopted. Dr. Ruml was

engaged as a consulting examiner and is still retained in that capacity.

Following the program adopted on the recommendation of Dr. Ruml, the examination for clerks and carriers in the Postal Service was revised after some experimentation and research. A small research unit headed by Mr. Guy Moffett, an experienced employee of the Commission, conducted this experiment. No radical change was made in the actual test, but by a rearrangement of the form and a revision of the arithmetic questions a considerable saving of time was effected in the rating of the papers, and the time of giving the examination was reduced by one-half. Revisions were made in a number of other examinations along similar lines. For example, changes were made in the arithmetic and geography of the railway postal clerk examination, effecting a material saving in the work of the examiners. An entirely new examination for bookkeepers was constructed. Wherever the type of examination would permit, the questions were so arranged that they could be answered in a single word and scored by means of a key or stencil. These methods were applied to many examinations from time to time, including those for junior scientific positions and for auditor of income-tax returns.

Experiments were undertaken on a new general clerical examination to take the place of the examination previously given for clerks engaged in miscellaneous clerical tasks requiring no previous training or specialized knowledge. Much study and experimentation were devoted to this examination before it was finally adopted and announced. It has been used for the field service in several of the thirteen civil-service districts and throughout the country for the departmental service in Washington. This test requires only half as much time to administer as the examination formerly given, and perhaps one-third of the time to correct, and it is believed to be a much more accurate measure of clerical abilities.

On July 1, 1922, an appropriation became available

290 Measurement in Employee Selection

especially for investigations looking toward the improvement of examinations. Since that time, Dr. L. J. O'Rourke, a psychologist who was a member of the Air Service Psychological Research Laboratory and of the Civilian Advisory Board of War Plans Division, has been in charge of the Commission's Research Section. He has done valuable work on the general clerical examination and is conducting a number of other research studies.³

Since its establishment the Research Division has conducted a number of studies looking toward improvement in the methods of selecting public employees. These have included studies of clerical tests, tests for post office workers, stenographer-typist tests, general adaptability tests for measuring general mental ability of applicants, oral examinations for investigators, and police tests.⁴

We shall examine the postal tests as illustrative of the type of measurement useful in selection of employees and as indicative of procedures essential to development of the instruments of measurement. The need for adequate means of selecting postal workers is obvious when we consider the numbers of applicants for positions and number of employees involved. Of all the departments of our federal government, the post office employs the largest number of workers. At the time the work was being done in the development of new examination methods for postal workers, the Civil Service Commission was examining between 60,000 and 80,000 applicants a year for positions of clerk and city carrier. It was for

³ Filer, Herbert A., and O'Rourke, L. J., "Progress in Civil Service Tests," *The Journal of Personnel Research*, Vol. 1, No. 11, March 1923, p. 484.

⁴ These studies are described in various reports of the Research Division contained in Annual Reports of the United States Civil Service Commission.

this large group of workers that the new methods of selection were studied and introduced, replacing old methods of academic examination.

On the basis of an extensive analysis of duties performed by the postal workers, several preliminary tests were constructed for trial. Each trial test was studied individually for its suitability for final inclusion in the new examination. Before any test was included in the final examination, it was determined that this particular test ⁵ —

1. Measures qualities which are essential for success in the work (validity).
2. Differentiates between degrees of ability.
 - (a) Is focused at the correct difficulty.
 - (b) Is of correct range of difficulty.
3. Is readily duplicated.
4. Is reliable: Is such that a competitor will make approximately the same grade on a later as on the original attempt.
5. Is objective.
 - (a) In administration.
 - (b) In scoring.
6. Is practicable.
 - (a) In time required and facility in administration.
 - (b) In time required and facility in scoring.
 - (c) In cost of printing.

The validity of the tests. The degree to which each test indicates the relative ability of those tested was determined by study of the relationship between the test scores and a criterion of efficiency established for trial groups of distributors already working in post offices. A

⁵ O'Rourke, L. J., "Report of the Director of Research," in *Forty-second Annual Report of the United States Civil Service Commission*, 1925, p. xliv.

combined criterion was used including (a) average number of pounds of first-class mail distributed by each employee during a period of six months, together with amount of time in minutes spent on this distribution; (b) records in a monthly case examination measuring the rate and accuracy with which each employee distributed mail into his distribution case; and (c) a subjective criterion based upon foremen's ratings of efficiency. The first test selected was the one having the highest relationship to the efficiency criterion. The next test selected was the one showing also a high relationship with the criterion and, in addition, measuring important qualifications not measured by the first. The third test selected was the one best measuring important qualifications not measured by the first two. The three tests finally selected included a general test measuring general ability requisite for the job and two special tests measuring more specific requisites for postal work. The nature of these is indicated by items reproduced from the "Sample Questions" furnished applicants (see page 296).

A final study to verify the value of the new examination, made on a group of mail distributors in the Chicago city post office, showed the relation between test scores and efficiency of the employees indicated in Fig. 20.⁶ The vertical line of the figure O-O is the dividing line on either side of which 50 per cent of the efficiency ratings of the whole group fall. Each horizontal bar represents one-quarter of the group tested. The top bar (A) represents the distributors who made the highest 25 per cent of the test scores; the lowest bar (D) those who made the lowest 25 per cent of the test scores. In each bar, the part to the right of the vertical line O-O in-

⁶ *Ibid.*, xlvii.

icates the percentage of that group above average in efficiency; the part to the left of O-O, the percentage of that group below average in efficiency. From this chart it may be judged that if a competitor makes a test score as high as the highest quarter tested in the trial group, the chances are 93 to 7 or 13 to 1 that he

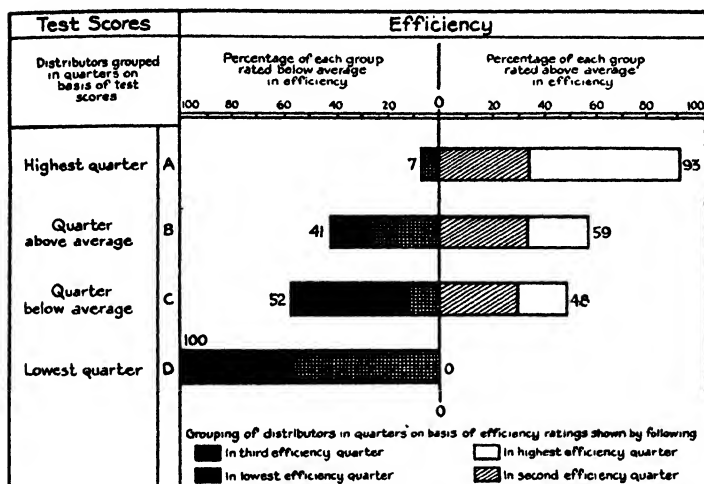


Fig. 20.—Relation Between Test Scores and Efficiency of Mail Distributors.

will be in efficiency above the average of those tested. On the other hand if he makes a score in the lowest quarter, the chances are 100 per cent that he will be below average in efficiency. Since it is probable that not more than 25 per cent of competitors in postal examinations will receive appointments (according to a statement in the Civil Service report), it can be expected that the new examination will greatly increase efficiency of employees. Fig. 21' shows this increase graphically.

[†] *Ibid.*, p. lxxiii.

294 Measurement in Employee Selection

The reliability of the tests. The new postal tests represent a marked improvement in reliability over the more subjectively graded tests which they replaced. The objectivity of the test forms insures equality of results of scoring by different examiners. O'Rourke emphasizes in the introduction of these new tests the importance of sample questions, supplied the competitor before he takes

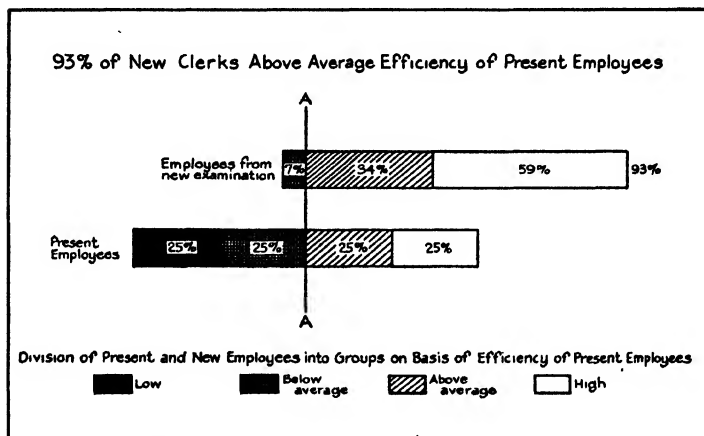


Fig. 21.—Improved Selection Made by New Postal Examination.

the examination, as an aid to reliability. Such a procedure reduces the unequal advantages which may result from the effect of practice and familiarity with test forms on the part of some who may have taken the examination before.

The difficulty of the tests. Scores made on the tests by the employees constituting the trial groups were studied to insure that the final examination be properly focused in difficulty. In the development of the general tests, each individual item or question was studied. Difficulty values were determined on the trial groups, and

items were selected so as to give a wide range of total scores. Items were included which were answered correctly by from 9 to 85 per cent of the trial group. For each series of the examination 100 items were selected. These were scaled in difficulty from the easiest to the hardest and were so selected that the difficulty increased throughout the test in equal steps, as on a scale.

Practicability of the tests. In the study of the newly devised postal tests considerable attention was paid to the problems of construction of future tests, administration of the tests, scoring of the tests, and cost of the new method. It might be conceivable that even increased validity of the new test would not justify its introduction at greatly increased time and expense for construction and administration. A statement from the original report on the tests indicates their advantage from the standpoint of these administrative problems.

The new clerk-carrier examination can be scored and handled in 40 to 50 per cent less time than the examination which it replaced. In an examination in which 60,000 to 80,000 compete, this means a saving of time required to score from 32,000 to 40,000 examinations in the old way. This will make it possible to make up the eligibility lists and certify eligibles in approximately half the time previously required to correct the examinations.⁸

These tests for selecting an important group of our federal employees have been discussed because they present the various problems which are met in the construction and evaluation of employment tests, and because adequate methods of meeting most of the problems were devised.

⁸ *Ibid.*, p. lxxi.

SAMPLE QUESTIONS—POST-OFFICE TESTS

GENERAL TESTS

Write the number of the *best* answer on the line at the right.

1. The business of mail-order firms has been greatly increased by the introduction of (1) special delivery (2) parcel post (3) postal savings (4) airplane mail (5) lock boxes -----
2. Letters are delivered promptly by the post office so that the (1) office can be closed on time (2) inclosures will not be lost (3) mail will not be heavy (4) letters will not be damaged (5) public may not be inconvenienced -----
3. A *fundamental* point is one that is (1) final (2) drastic (3) emphasized (4) essential (5) difficult -----
4. The saying, "To do, one must be doing," means most nearly (1) What you do, do thoroughly. (2) More is needed than good intentions. (3) Think before you act. (4) By our deeds we are known. (5) Well begun is half done. -----

In the sentence below, the word printed in *italics* has been misspelled. It is spelled according to its sound. Write the correct spelling of this word on the line at the right.

5. The plan was *sankshunned* by the committee -----
6. Which one of these five may be applied to both *books* and *magazines*, but not to *postman*? (1) expected (2) reliable (3) accurate (4) authorized (5) published ... -----

In the following question, the first two words in capital letters go together in some way. Find how they are related. Then write a *number* to show which of the last five words goes with the third word in capital letters in the same way that the second word in capital letters goes with the first.

7. SACK is to MAIL as PURSE is to (1) money (2) suitcase (3) bag (4) owner (5) luxury -----
8. Over what body does the Vice President preside? (1) Senate (2) House of Representatives (3) Interior Department (4) Supreme Court (5) Cabinet. -----

Answer the question from the quotation which follows it.

9. What one word in the quotation describes the person who is not interested in public welfare? -----

"In almost every community there are certain men and women who are known as public-spirited. Others may be selfish and act only as their private interests seem to require."

10. If 4 men can distribute 700 letters in 2 hours, in how many hours would they distribute 1,750 letters, at the same rate?.... -----

SORTING

In the SORTING SCHEME below, each square represents a box for mail going to the cities named in that square. You will be required to study the SORTING SCHEME and then write after each city in the following list the *number* of the box in which you would put mail for that place. Look at the first name in the list, "Harbur." The number "2" is written after it because Harbur is in the box numbered 2. "Leadwood" is in box number 8, so "8" should always be written after Leadwood.

298 Measurement in Employee Selection

Work straight down each column, taking the cities in order. You will receive no credit if you skip cities and scatter your answers.

Study the SORTING SCHEME for 10 minutes, to get it thoroughly in mind before beginning to write.

SORTING SCHEME

Red Bank 1 Painter Carter	Denver 3 Rayburn Sunset	Texan 5 Mesa Grande	Edison 7 Milbrook Appleton
Harbur 2 Refuge Concord	Eastlake 4 Boston Lakeview	Randall 6 Lowell Porter	Leadwood 8 Fox Morton
	Wheeler 9 Forest Sumter	Camden 10 Roswell Chester	

You may look back at the SORTING SCHEME as often as you wish.

You may not have time to finish the test. Do as much as you can in the time allowed.

<i>City</i>	<i>Box No.</i>	<i>City</i>	<i>Box No.</i>
Harbur	<u>2</u>	Red Bank	----
Leadwood	<u>8</u>	Lowell	----
Fox	----	Carter	----
Edison	----	Denver	----
Porter	----	Sunset	----
Eastlake	----	Edison	----
Grande	----	Morton	----
Painter	----	Porter	----
Milbrook	----	Denver	----
Boston	----	Lakeview	----
Camden	----	Mesa	----
Milbrook	----	Appleton	----
Grande	----	Forest	----
Randall	----	Chester	----
Wheeler	----	Texan	----

<i>City</i>	<i>Box No.</i>	<i>City</i>	<i>Box No.</i>
Refuge	----	Concord	----
Boston	----	Rayburn	----
Painter	----	Lakeview	----
Roswell	----	Morton	----
Sumter	----	Eastlake	----
Refuge	----	Appleton	----
Fox	----	Texan	----
Lowell	----	Boston	----
Concord	----	Sumter	----
Randall	----	Painter	----
Roswell	----	Carter	----
Red Bank	----	Wheeler	----
Mesa	----	Edison	----
Refuge	----	Rayburn	----
Fox	----	Leadwood	----

FOLLOWING INSTRUCTIONS

This is a test of your ability to follow instructions. All directions must be followed exactly as shown in this sample test.

Below, at the left, is a list of post offices, called a SORTING SCHEME. After each of these offices is a letter. For example, after "Bowers" is the letter "A". This refers to the "A" in the KEY at the right, which reads "A Felton 4." The "A" after "Bowers" means that mail for Bowers is routed by way of Felton.

The numbers after the names in the KEY indicate the trains on which mail for those post offices must be placed. After "Felton" in the KEY you will find the number 4. This means that mail for Felton is sent on Train 4. Since mail for Bowers is routed by way of Felton, mail for Bowers, also, would be sent on Train 4.

SORTING SCHEME

AllenC
 Bowers ...A
 Camden ..C
 Daly ...I
 Denham ..E
 DoverC
 FeltonA.....

KEY

Mail sent by way of—
 A Felton 4
 B Union ...8
 C Camden .6
 D Woods
 E Allen
 H Turner ...9

300 Measurement in Employee Selection

SORTING SCHEME		KEY	
Malter	.. D.....	I	Dover
Turner	...H.....	
Viola	... B.....	
Woods	... A
Union	...B.....	
.....
.....

YOU MUST FOLLOW DIRECTIONS EXACTLY AS GIVEN.
Make your numbers and letters clear, to avoid mistakes.

Look at the name "Woods" in the KEY. It is not followed by a number. Write after it the *letter* which you find after "Woods" in the SORTING SCHEME. Your KEY will now read "D Woods A." Find the letters after Allen and Dover in the SORTING SCHEME and write them after those names in the KEY.

Never put numbers in the SORTING SCHEME.

On the line after each of the following offices, write the *number* of the train on which you would send mail for that office.

To find the number which should be written after Viola, look for Viola in the SORTING SCHEME. After it is the letter B. This refers to KEY B Union 8, and means that mail for Viola is routed through Union on Train 8.

After Denham is the letter E. This refers to KEY E Allen C, and means that mail for Denham is routed through Allen by way of C, and KEY C reads Camden, on Train "6." Write "6" after "Denham" in the list below. Now write the train numbers after the others.

Viola	... <u>8</u>	Bowers	Turner
Denham	Daly	Malter

You now receive Bulletin No. 1:

CHANGES IN ROUTING

Never change the letter before the name in the KEY. When a letter or number is changed, it is always the letter or number *after* the name.

Measurement in Employee Selection . 301

(Make changes in both SORTING SCHEME and KEY if the names are in both.)

Woods by way of C

Dover by way of B

To make the change for Woods, cross out the "A" after Woods in the SORTING SCHEME and write "C." Then your SORTING SCHEME for Woods should read: "Woods ~~A~~ C." This means that mail for Woods is now sent by way of "C Camden 6." Next look for "Woods" in the KEY, and change the "A" after it to "C." Make the change for Dover so that your SORTING SCHEME will read: "Dover ~~C~~ B," and the KEY will read: "I Dover ~~C~~ B."

After making the above changes, write the *number* of the train on which you would send mail for each of the following offices:

Woods . . ----	Camden . ----	Dover . ----
Felton . . ----	Union . ----	Allen . ----

Next you receive Bulletin No. 2:

CHANGES IN ROUTING

(Make changes in both SORTING SCHEME and KEY if the names are in both.)

Change KEY C to read: C Camden 2

Change KEY A to read: A Train 5

Felton by way of B

Change KEY E to read: E Allen 7

Allen by way of E

To make the change for KEY C, cross out the 6 after Camden in the KEY and write "2," because the train for Camden has been changed from "6" to "2."

To change KEY A, cross out "Felton 4," and write "*Train 5.*" This means that mail for offices marked "A" is no longer sent through Felton but is routed direct on Train 5.

302 . Measurement in Employee Selection

Write the *number* of the train on which you would send mail for:

Bowers .. ----	Felton .. ----	Daly .. ----
Dover .. ----	Denham .. ----	Allen .. ----

CHAPTER XVIII

Psychological Measurement in the Control of Employees

ONE of our most outstanding industrial psychologists, urging engineers and industrial managers to approach their problems in a mood of exploration, recently stated, "But we must not leave behind our conceptions of science, for without them we cannot go far. One concept indispensable for our purpose is that of *measurement*. Another is *correlation*. These are basic to all others, for scientific research is the precise determination of relationships between variables."

The variables which the industrial manager should be interested in measuring, and which the psychologist can help him to measure, include the aspects of a person's behavior which are significant in his work and the factors or conditions which determine the worker's output and produce the satisfactions and happiness which he should have.

Bingham has outlined these aspects of behavior and factors influencing behavior according to the accompanying table:

1. The tempo at which he chooses to work.
2. The mechanically imposed tempo to which the workman can adjust himself; or the number of machine parts (e.g., spindles) or machines (e.g., looms) which he is able to operate.
3. Avoidance of errors, spoiled work, and accidents.

4. Speed and efficiency in eliminating disturbances and irregularities in machine or materials.
5. Quality of product, in the narrower sense.

Let us stop for a moment to consider which of the instruments of measurement in possession of the psychologist may help to solve some of these problems. A peek into a modern psychological laboratory shows us tachistoscopes, ergographs, chronoscopes, instruments for measuring reaction time, sphygmomanometers, and metabolism machines.

Perhaps the worker is too slow in the performance of his job. A micro motion camera and time and motion studies show that there is too much wasted motion in his method of doing the task. His efficiency is measured in terms of reductions of waste motions. Perhaps another worker always feels tired or worn out at the end of his day's work. A metabolism machine will measure his lack of energy production and will tell whether he needs a bit of thyroid to speed up. Still another worker may get irritated at only slight annoyances or "blows up" when something goes wrong. A blood-pressure instrument may reveal that this worker has a definite physical disturbance which is the basis for his apparently neurotic condition. Another worker may be particularly prone to accidents in the operation of his machine or may find that he is unable to coördinate his own movements with the speed of the machine. An instrument for measuring his reaction time may reveal that his movements are naturally too slow for the type of work he is trying to perform.

In addition to these mechanical precision instruments, the psychological laboratory utilizes tests of various general and special mental traits to measure fitness of the worker for the job; rating scales which make as ob-

jective as possible supervisory ratings of personal efficiency; learning curves which show rates of progress to be expected in teaching employees; and diurnal curves of efficiency which show variations in output to be expected for a day of the given length. Finally psychology is interested in environmental conditions in which the various factors have been measured, many of the things in which the industrial manager is interested. Such may be mentioned as studies of distractions in which not only the effect on output has been measured, but also the cost in energy consumption or effort; and studies of the effect on behavior of external environmental factors such as light, weather, and season, of best routes or organization for tasks, and of habit influences when several tasks are performed.

There are, of course, many problems for which psychological measurement has no answer. There are needed in many of these instances more precise units, more accurate scales with which to measure human traits and with which to detect changes in various aspects of human reactions, but more use could be made of the mental and physical yardsticks at hand. Even the crudest of them are better than trusting to personal impressions, ordinary common sense, or intuitive judgments.

To show some of the contributions of psychological measurement in industrial management we shall discuss three examples: first, an example in accident prevention in a public utilities company; second, an example of improvement in efficiency through time and motion study; and third, an example of the use of psychological measurement in management in a large department store. These three examples have been selected because they show variation in the types of measuring devices

used, and variation in the types of problem which may be met.

I. Example of Measurement in Relation to Accident Prevention

The organization which we shall use as an example here is the Boston Elevated Railway. The study discussed was one conducted under the direction of Bingham and begun in 1926. At the time the study was started, the problem that confronted the management was the high cost of accidents. The railway safety record was fairly high, but a million dollars a year was being expended for personal injuries, property damage, and claims. No radical improvement at that time was to be expected from better methods of hiring new employees, because there was almost no labor turnover. The problem lay with the experienced operators, in fact with a relatively small percentage of these, for it was found on study that about 20 per cent of the operators caused practically all of the accidents. In other words, there were about 20 per cent who might be termed "accident-prone."

Bingham began his study by first making a job analysis and by collecting all the useful data which might already be at hand. When he began work on the problem he found already available the individual records of accidents in full detail, personal history information which the employees had supplied on their application blanks, medical examinations of the employees made at the time of hiring, records of earnings, overtime, absences, errors in making out cards, errors in turning in cash receipts, disregard of rules, and a long list of other things which one might expect to have available in any electric railway company. Bingham set about the problem with the

object of finding whether any of these data were related to accidents; in other words, could any of these data which were available serve as a human yardstick so far as proneness to accidents was concerned.

He did find, among the data which he had available, some which were apparently related to the tendency to accidents. Among other things he found that each man had a "coasting record." These records represented the time during each run when neither brakes nor power were applied to the car. The motormen were instructed to coast when possible to conserve wear and tear on the equipment and to economize on power. Each man's record was computed and posted periodically. Bingham suspected that these records might reveal those who were most competent and conscientious and should consequently have fewer accidents. This proved to be true. The relationship between these efficiency records and accidents was not close, but it was real.

One more example of an objective measure which Bingham found to have a positive relationship to accident susceptibility was blood pressure. Poor health as shown by the blood-pressure records was evidently a causal factor in accidents. Bingham's method was to go through the list of measurements which he had available, to correlate each with accident occurrence, and to find which ones had any relationship—low or high—to accidents.

After a study of this nature by which he was able to discover some of the important factors to be measured, Bingham set about actual work with individual motormen. The men selected for first attention were those whose previous records showed them to have had the largest number of accidents. In each instance the details of the accidents were fully studied for clues as to prob-

able causes. In each of these cases one of the inspectors actually observed the man as he performed his job. The operator's whole record was studied for indications of any divergences from the usual attitude toward his job. Finally, the man involved in an accident was interviewed. If necessary a new medical examination was made. With all these facts and all these measurements at hand, a diagnosis was made, an individual one for each case, and a definite plan of cure prepared. In most of the cases the problem was solved in such a way that these men could be kept on the same kind of work. A majority of the causal factors of the accidents occurring in this company were factors which could be eliminated provided the accident-prone employees were sufficiently studied by the definite means of measurement at hand. We may summarize Bingham's work on this particular problem by quoting two paragraphs from his report: ¹

It so happened that a number of these accident-prone men were at the time operating on the Harvard-Dudley line on Massachusetts Avenue. This is one of the most difficult routes in the Boston area. It crosses many main thoroughfares. The cars are huge and fast. Up to November, 1926, no month had gone by with fewer than 48 accidents on this route. But in December, the next month after this work with the accident-prone men began, there were 23. As a consequence, the following year the road as a whole achieved a 17% reduction in collision accidents. A similar further reduction was made in 1928; for this achievement, the Railway was awarded the Anthony N. Brady Memorial Gold Medal in the national competition among street railways for the best record in accident reduction.

Of all the kinds of factors that proved to have significant relationship to a motorman's actual perform-

¹Bingham, W. V., "The Science of Work," *Technology Review*, July 1932, p. 374.

ance in avoiding accidents, those which showed the closest relationship were matters of physique, health, and eye-sight. That was to be expected. Almost equally close were the factors of mental attitude toward the responsibilities of the job. Next in importance were habits of operation and items of knowledge about the job—specifics which can be taught. Then came the factors which may be grouped under the heading of aptitudes for this kind of work. In other words, we found that physique and health, right attitude, knowledge and skill, and natural aptitude for work of this character, are all factors in the equation.

II. Time and Motion Study

For most industrial operations there are several ways of performing the job. There are varying numbers of motions that may be made, and there are various orders in which given motions may be made. We cannot be sure without study whether all the motions a worker makes are necessary or whether the order in which he makes them is the best order. The worker himself cannot tell whether he is doing the job in the best way, and even an expert watching him cannot be sure of his judgment. Furthermore, the speed at which motions are ordinarily made is such that it is difficult to observe them and to get an accurate idea of their order. For these reasons it is desirable to have a method of measuring these motions—a permanent record—one which can be studied at leisure.

A technique which has been widely used for this is the continuous photograph method. A moving film is used to record the motions as the worker performs his task. Since in most instances only the motion of a small part of the body (such as the motion of the hand) is of interest at one time, a small battery lamp is usually mounted on the part to be studied. The camera with the shutter

open is then placed in front of the person. The figure of the individual may be blurred, but the lamp gives a continuous streak of light on the film which constitutes a permanent record of the motion of the hand or other part of the body. To indicate the time duration of each motion, a time record is made on the film. In order to study the third dimension or depth of the motions, pictures are usually taken from more than one angle. Motion models in wire are sometimes constructed from the various photographs of a motion.

Motion Study in Bricklaying. The classic example of increased efficiency produced by motion study is Gilbreth's work with bricklayers. For hundreds of years bricklayers have been lowering and raising the upper portion of the body with every brick they laid. It developed as a result of careful investigation that the bricklayer was using something like eighteen different motions, many of which were unnecessary. For instance, he picked up a brick with his left hand and turned it around until he found the best face of it to place toward the surface of the wall. Then he reached for the mortar with his right hand and, after spreading the mortar, put the brick in place on the wall. The workers were trained to reach for the brick and mortar simultaneously. There was no particular point in leaving the right hand idle while reaching for the brick with the left. These two operations could just as well be done at the same time, because neither one required accurate coördination and they could perfectly well be done automatically. Then, to eliminate the inspection of each brick to determine its best face, it was arranged to have this done by an unskilled worker who arranged the bricks in a "packet" with the best face always in the same direction. The bricklayer then did not have to make any discrimination on that point. The bricks were supplied, furthermore, at waist height, so that it was not necessary to stoop every time. The men were

taught not to waste time picking up any mortar that they happened to drop. By this and other devices, the initial eighteen motions were reduced to about five, and the average worker, instead of laying 120 bricks per hour, was enabled to lay 350. This is typical of one of the earlier motion studies with rather striking results.²

A great many other industrial operations have been studied by the motion study method since its introduction. Such study has in every instance led to improvement in working efficiency. Mere mention of some of the operations studied will indicate the extensive applicability of the method. There have been studies of the motions involved in folding handkerchiefs, dipping chocolate-coated candies, packing candy, metal polishing, typewriting, assembling carburetors, packing potatoes, coal mining, and housework in the kitchen.

III. Measurement in Department Store Management

As an example of the use of psychological measurements in industrial management in the field of department store work, we quote from the work of V. V. Anderson at Macy's department store in New York City. The reader will note the many yardsticks for human abilities employed by Dr. Anderson in his study of the "problem cases."³

About 20 per cent of the employees of mercantile establishments—and this is probably true of other business and industrial organizations—are what may be called "problem" individuals. That is, as personnel material, they are either liabilities or potential liabilities to the business man. It is from this group that are

² Burt, Harold Ernest, *Psychology and Industrial Efficiency*, Appleton-Century Co., New York, 1929, pp. 95-6.

³ Anderson, V. V., *Psychiatry in Industry*, Harper & Bros., New York, 1929, pp. 8, 10, 12, 18, and 25.

drawn the repeated transfers from job to job, or resignations, or lay-offs. These are the work failures that in the majority of cases are a drag on any organization.

As the employer sees these individuals they are production problems, or chronic health problems, or chronic attendance problems, or serious attitude problems, or marked disciplinary cases, and the like, and from his viewpoint he is better off without them.

As the psychologist sees them, they all present distinguishing physical and mental characteristics that underlie and explain not only their job maladjustments, but faulty adjustive efforts and failures in other life situations. In studying the life histories of these individuals, and analyzing their careers, he is impressed with three outstanding causative factors commonly underlying work failure—(a) a maladjusted personality; (b) specialized job disabilities; (c) faulty physical conditions.

Of course, such cases can be dumped wholesale into the community to shift for themselves. This might, after all, be a simple and easy way of solving the entire problem, for it is common enough for the Management to say, "We are not running a charitable institution; why should we feel any responsibility in this matter?"

But there is another viewpoint—that many of these individuals may, through proper study and treatment, be adjusted and become assets to the employer, thus cutting down turnover among employees and increasing production efficiency. A few case summaries presented in this study demonstrate what can be accomplished in this direction, and as our statistics show, a sufficiently large number of problem cases will improve under treatment to make the application of such measures profitable—not only in terms of human salvage—but in terms of dollars and cents.

The methods employed in conducting a complete study can, for practical purposes, be roughly classified into the social history, the job behavior study, and the physical and mental examination. These four fields go together to make up a thorough-going picture of the

whole individual—his behavior toward his work and toward major life situations, as well as those influences—constitutional, or home, or school, or work—that have contributed most to his career. . . .

The psychological method involves the objective and quantitative evaluation of certain measurable mental qualities. It is of enormous aid to the diagnostician, in that it furnishes him accurate and concrete evidence of the capacities of the individual, and enables him to compare the performances of individuals with each other, and with standards obtained from careful studies of large groups.

Illustrative Case. Female, aged 24 years.

Problem: Referred by department head with the following statement: "She should be tested, as she is dull and stupid. I think she is poor material, and I have already made out a lay-off slip and sent it through to the general manager's office, but I should be glad to have you study her."

Physical Findings: Height, 5 feet 3 inches. Weight 117 pounds. Physical examination disclosed some inflammation of the gums. She is in need of dental work. Her vision is defective and she should wear glasses. She has lost twenty-four pounds in the last four months and at the present time complains of loss of appetite, fatigue, and "needing a rest." She has had a good many consultations with the hospital physicians recently for various minor complaints—"headaches, indigestion, insomnia, nervousness, run-down." Physical examination does not disclose any positive findings.

Mental Findings—Psychological study: Individual secures an intelligence quotient of 88, which gives her a classification of dull average intelligence. She is somewhat slow in performing her psychological tests, but is accurate and painstaking. Her attitude was particularly good toward the examination.

Personality study: She has a fair sales appearance. Is pleasant, agreeable, seems most anxious to please, is neat, fairly well dressed, but is not attractive or good-

looking. She wants to succeed at her work but is over anxious and fearful. She is naive and too shy and retiring to be convincing. She has suffered much from insomnia, and periods of acute depression—worrying over her home troubles; is lonesome and homesick, yet realizing that she would not be happy to return home under present home conditions. She is too shut-in to make friends easily. So since coming to this country from Great Britain she has established no friendships. As a consequence, she feels out of place, disheartened, and cries a great deal. Of late, there has developed a great fear of losing her job, and inasmuch as she is totally dependent upon her weekly salary, she does not know what will become of her in case she does lose it. She does not seem to know how to make headway, and appears to be bewildered by the size and complexity of things in a great store—having had no experience in the past that would teach her how to meet such problems.

She is responsive, polite, and courteous, but obviously has a great deal of pessimistic reverie, with a marked lack of self-confidence. She becomes greatly frightened when admonished or criticized by her superiors, and has frequently to leave the floor to go the girls' room.

Job behavior: Her contact with customers is good, though not particularly effective. They like her because she is anxious to please. However, on the floor and at her work she seems detached a great deal of the time, day dreaming, mentally preoccupied, and slow, but when awakened out of her reverie she is responsive and obliging. She seems timid in her relations with the other girls, and does not appear to know how to make herself easy and natural with them. Her reaction to her superiors is more that of a frightened child. Her knowledge of stock is inadequate.

Social Findings: Her mother is dead and her father married the second time. The stepmother has been very unkind to her. Her only brother was killed in the war, and her two sisters are married. She left home because of unhappy home relations, and determined to

make her way in this country. Her education was of grammar-school character, but she has had several years' experience in selling in a store in a small town.

Conclusions and Remarks: We do not recommend lay-off in this case, but do advise adjustment work. We have here a young woman of good enough intelligence and fair enough sales appearance, who—providing we can secure an adjustment of her personality disorders—will make an acceptable sales clerk, inasmuch as she has a very responsible and dependable nature, is courteous and anxious to serve, and will give her best to the department. She is in need of a physical and mental hygiene regime. Her weight needs to be built up, her insomnia overcome, and her depression cleared. It is believed that if this can be done her detached attitude and obsessive reveries which seem to be the big factor in handicapping her sales ability will disappear.

Follow-up: After the report was made, the buyer took a different attitude toward this employee. She became more friendly and understanding in her relations with her. She talked with her and sought to be helpful. Realizing that much of the problem was shyness and timidity, based upon a strong inferiority background, she made a different approach to her and sought to aid and guide her rather than to intimidate and frighten her when she made errors. She kept in close touch with the psychiatrist, who sought to give her an insight into her personal problems, and who worked out with her a more satisfactory way of meeting them. Her insomnia disappeared altogether within the first week. She regained her appetite, began to increase in weight and improve in her attitude toward herself and the difficulties in her job. Self-confidence increased. People seemed to be interested in her. The world took on a brighter outlook. Her sales record improved. She got acquainted with some of the girls—went out to luncheon with them. Began to drop into the rest rooms at the lunch hour, and took some part in the recreation of the other employees. She kept a diary, and we saw

this note: "I am doing the best I can. My weekly quota is \$250. My books show that I have made \$289 this week. I am much better now and feel a little happy at times. I know what it is to be hungry, and sleep good too."

The psychiatric worker's report from the buyer, two weeks after she was referred to us for study, is as follows: "She does anything you ask her to do. Is always cheerful about it. I would like to keep her. She is slow at getting things, but once she has them, then they're there." Later she states: "Very favorable. Special emphasis on willingness and good nature with customers."

She has been closely followed up, and the last report from her department is that she is doing well and they do not want to lose her. She became, in the long run, a well-adapted and satisfactory employee. . . .

Illustrative Case. Female, aged 19 years. Cashier.

Problem: Referred by superintendent of department, on account of the great number of errors and shorts which she made, with the statement that if she did not improve soon, she would be laid off.

Physical Findings: Height, 5 feet 8 inches. Weight, 110 pounds. Considerably undernourished. Has had several attacks of tonsilitis and recently has had a growth on plantar surface of right foot, which has been removed. On recent examination in our hospital, she has been rated fair in physical condition except for underweight.

Mental Findings: She has an intelligence quotient of 88, rating her as dull average in intelligence. She was found to be slow in speed tests, fair in learning ability, poor in motor dexterity, and showed a tendency to error-making in accuracy tests. She has a pleasant and agreeable personality make-up, though a little peculiar and not very accessible to a personal interview, but does not concentrate well on the things at hand, showing a definite tendency to mental reverie. She has

a good general appearance, is rather attractive, neat, and well dressed. There is considerable emotional upset. She is rated poor on our cashiering test, and we do not believe she should be kept at this work. She has a good personality for sales, and if her mental conflicts can be successfully dealt with, she ought to make a good sales clerk.

Social Findings: Mother dead; father married second time. Home situation unhappy. Had to leave home and is boarding with a private family. Said that she was forced out of her own home. Has to support herself. Has had two years of high school. This is her first job.

Conclusions: This girl will never make a successful cashier. She is too slow in motor speed, her dexterity is poor and is not accurate by nature. She should succeed fairly well at sales. Transfer to sales is recommended.

Follow-up: Girl was transferred to house furnishings as sales clerk, where she did well, being rated as satisfactory by the head of the department, who expressed himself as pleased with the new clerk several months later. Inasmuch as she did not like it in the basement, she was transferred, after a period, to a department on the first floor, where she did fairly well. She has been under the supervision of the psychiatrist, and has improved considerably in her personality disorder. Her health is excellent, her weight has improved under a careful dietary regime, and her sales are fairly good. The average selling cost of her department is 4.35 per cent, while this employee's selling cost is 4.29 per cent, which means she is under the average cost of her department. Her interest in stock has been marked. She has proved reliable in stock work and gradually has taken on responsibility. Her contact with other girls has been good and she is now being tried out for head of stock—a promotion for which there is considerable competition. The lesson to be learned in this case is one of suitable placement according to the ability and personality make-up of the employee.

318 Measurement in Employee Control

In concluding this chapter, we again quote from Bingham his closing remarks to industrial explorers, as follows:

We have skirted along the margin of territory as yet almost unexplored. Some of those who have heretofore prospected in these regions have already struck pay dirt. The means and instruments for digging out the treasure and refining it are at hand. We have our ideals of measurement, and the techniques for finding out how the variables measured are related to each other. Some of the ore does not even need to be dug out, because it has already been mined by time-study men, and wage clerks.

Many of you may be supervisors, foremen, or superintendents, face to face with practical problems in human engineering. If you are, don't overlook these data about individual differences among workers. Remember your ideals and techniques of scientific method. As engineers you deal with machines, materials, money, and men. In the past, too many engineers have made the mistake of utilizing science only when dealing with machines or materials or money. Men, too, can be measured.⁴

⁴ Bingham, W. V., "The Science of Work," *Technology Review*, July 1932, p. 393.

PART VI

MEASUREMENT OF THE MORE GENERAL
TRAITS OF PERSONALITY



CHAPTER XIX

The Nature of Measurement of Personality

WE USUALLY think of an individual's personality as expressive of his effectiveness in social relationships. This aspect of his make-up seems the all-important one in determining how well he will get along with his fellow beings; in deciding whether he will be aggressive enough to make the best utilization of his abilities and capacities; in determining to what extent he will impress and influence others; and in determining how he will meet the situation in an indefinite number of other ways. Allport has defined personality as *the individual's characteristic reactions to social stimuli and his adaptation to the social features of his environment*. He has analyzed these reactions and has included as the factors of prime importance to one's personality: (1) *intelligence*, or general adaptability; (2) *motility*, or speed of one's reactions; (3) *temperament*, or one's emotions, moods, and attitudes; (4) *self-expression*, or one's manner of personal adjustment in a social world; and (5) *sociality*, or one's social participation and manner of one's solution of social problems. The term *personality testing* has generally been limited to the last three of these, since there are more or less distinct problems of quantitative measurement setting off these three from the more easily measured first two. "Personality tests," then, include tests for a large number of non-intellectual

traits rather loosely grouped under the heading of personality.

The work in the field of personality measurements has grown largely out of testing in the more objective fields of mental ability and achievement. The many instances of inability to predict with accuracy school performance, vocational success, or other behavior on the basis of measurements of intellect and knowledge, have emphasized the need for quantitative measurements of personal attributes which enter into performance or success. Because of their success with quantitative measurement of intelligence and various aspects of achievement, psychologists have been encouraged to extend their attempts and to try similar methods for measuring the more general personality traits.

I. The Difficulties of Measuring Personality Traits

The quantitative estimating of personality traits, however, has lagged behind the objective measurement of mental ability and school achievement. Even today we must still regard measurement of personality traits as in an early experimental stage. There are several reasons for the lag, among which may be mentioned the following:

(a) *Complexity of the thing to be measured.* There is no denying that personality is a complex entity, and that its complexity makes its measurement more difficult than the measurement of some simple part of our make-up. And yet, we might say, complexity does not prevent the quantitative study of human traits. Intelligence is measured with more than a fair degree of accuracy and it is certainly a complex quality. The difference is that psychologists and those who utilize measurements of in-

telligence have come to some general agreement as to what general mental ability means and what part intellectual traits play in behavior. There is no such agreement as to personality and personality traits. Perhaps such an agreement can never be reached. Perhaps the separate personality traits do not summate into a whole concept of personality in the way that intellectual traits seem to summate into general intelligence. Or perhaps it is that the various separate elements of general intelligence are so highly correlated that sampling a few gives an indication of the whole; whereas the various elements of personality are not so correlated that a moderate sampling of the separate elements can give a valuable total measurement.

(b) *Difficulty in setting up problems and tasks.* It is difficult to set up means of arriving at satisfactory measures of a personality trait once it has been defined. It is quite feasible without great difficulty to set up sample tasks in which the extent of a person's information about a given subject is measured, or his ability to comprehend printed material of a given difficulty is estimated. But it is another matter to set up sample tasks by which a person's aggressiveness, or his leadership ability, or his morality will be measured. Even if we set up and standardize sample situations for a test, the very fact of their experimental nature may defeat our whole purpose. If, as has often been attempted, our set-up is a problem situation on paper, the situation is hypothetical instead of actual, and the answer of the individual taking the test is likely to be hypothetical instead of actual also. Social behavior is to a large extent a matter of habit; a person may know the correct thing to do and so express it on a pencil-and-paper test, but fail to act accordingly in the actual situation.

(c) *Difficulty in evaluating responses to test questions and problems.* Establishing standards as to correctness of responses to social problems is not so easy as establishing standards of correctness as to meaning of vocabulary words, as to the solution of arithmetic problems, or as to memory of something one has just read. Accuracy of measurement of personality traits suffers from the necessity of depending, in too many instances, upon subjective judgments in evaluating the responses to test problems.

(d) *Lack of criteria for evaluating our methods of measuring.* No test or instrument of measurement for human traits can be used with confidence until its validity has been demonstrated. Investigations into the validity of measurements of personality traits almost always meet a great obstacle in the difficulty of securing independent measures of the traits to be studied. We may have an objective test that purports to measure one's ability to make judgments in social situations, but how are we to demonstrate that the test actually measures such ability? Where are we to secure a measure of such ability, independent of our test, with which to check the test results? No objective production record of social judgments on a large number of persons is available. No one form of behavior or social participation can be picked out that depends upon ability to make judgments in social situations alone. If we utilize the estimates of others as to the ability of our test subjects to make social judgments, the estimates are subjective opinions and are almost certain to be somewhat unreliable. At best, the validation of our test is likely to be somewhat inconclusive.

When all of these difficulties and problems have been pointed out, the reader should perhaps be warned against too dark an outlook toward the quantitative study of

personality traits. Much has been accomplished, and with a knowledge of the limitations of our present measuring instruments, much of value can be obtained by them.

II. Methods of Personality Testing

Three methods have been commonly used in the construction of devices for measuring personality traits: (1) the *Rating Scale*, (2) the *Questionnaire*, and (3) the *Objective Test*. The last of these we have met often in our discussion of mental and achievement tests. The other two are more common in the field of personality testing because of the difficulties of adapting the objective test to this type of measurement.

1. **The rating scale.** The rating scale furnishes a means of securing and recording quantitative estimates of the amount or degree to which an individual possesses the traits to be rated. The ratings represent subjective impressions of some judge. Usually the judge is another person who is in position to know the individual being rated, although "self-ratings" are sometimes made. There have been various types of rating scales developed, among which are: (a) the Man-to-Man Rating Scale; (b) various types of Numerical Rating Scales; (c) the Descriptive Rating Scale; and (d) the Graphic Rating Scale.

(a) *The Man-to-Man Rating Scale* is of importance primarily from a historical standpoint, being seldom used now. It was developed during the World War and was widely used in personal ratings made in the Army. In accordance with this method, the one who is to rate a group of individuals in a certain trait first sets up a scale of variation in the trait, each degree in the scale being represented by a man recognized to possess that degree

of the trait. For example, in assigning ratings of a trait in the Army, the rating officer was furnished with a blank containing the definition of the trait and an outline as follows:

Highest	15
High	12
Middle	9
Low	6
Lowest	3

On the first blank line after the word "highest," the rating officer wrote in the name of a man who, in his mind, best represented the highest amount of the trait to be rated. In a similar fashion, he selected a man well known to him who represented each of the degrees. Then those whom the officer had to rate were rated by direct comparison with the key men. This man-to-man scale has the advantages of being concrete and easily understood. It has the disadvantages of being cumbersome, of consuming too much time in its construction and application, of allowing for too great discrepancies due to differences in selection of key men by different raters.

(b) *Numerical Rating Scales* are those in which percentage or point values are assigned as quantitative estimates of the degree of possession of a trait. For example, a judge may rate an individual in coöperativeness by assigning a rating of 100 per cent, 90 per cent, or 80 per cent for relatively high degrees of coöperativeness; and 30 per cent, 20 per cent, or 10 per cent for low degrees. The scale may have a range from 0 to 10 or from 0 to 6, or any other numerical range instead of being in percentage terms. Such a method of rating is simple, but unless used in conjunction with another method usually fails to yield accurate enough results. Such

methods rarely yield comparable enough results from one rater to another, since different raters seldom agree upon just how much 50 per cent, or 5 points, represents.

(c) *The Descriptive Rating Scale* consists essentially of a collection of descriptive phrases; those are checked that best describe the characteristics or traits of the individual being rated. The descriptive phrases may be arranged under certain traits, the four or five descriptions representing the degrees of the trait; or there may be no such arrangement. The former arrangement is illustrated by the following phrases representing various degrees of a certain aspect of social behavior:

1. Extremely breezy and informal.
2. Cordial and congenial.
3. Meets one half-way.
4. Somewhat reserved.
5. Constrained and formal.

The Probst Rating Scale¹ for employees is an example of the latter type. Items 15 to 20 in this scale of about 100 items illustrate the arrangement:

15. Might often be more considerate.
16. Usually pleasant and cheerful.
17. Always courteous.
18. Cranky disposition.
19. Often seems dissatisfied.
20. Often grumbling or complaining.

Quantitative measurements are usually obtained by assigning numerical values to the various phrases which may be checked. The rating itself, however, should be given without reference to these numerical assignments. The descriptive scales have the advantages of being

¹ See Chapter XIV for discussion of this scale.

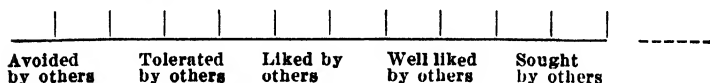
328 The Measurement of Personality

fairly concrete and of avoiding many of the difficulties of the numerical scales.

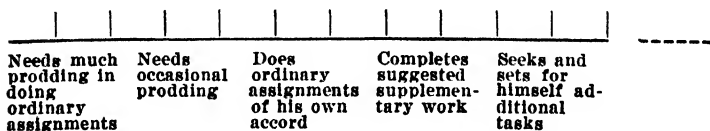
(d) In *Graphic Rating Scales* each trait to be rated is represented by a straight line signifying the range of degree of possession of the trait. One end of the line represents the least amount of the trait; the other end, the greatest amount. A person is rated by making a check along this line at a point which corresponds to his degree of possession of the trait being rated. The graphic rating scale is usually combined with the descriptive type, short descriptions of the degrees of the trait being placed under the rating line. Numerical scoring is done by assigning values to the different divisions on the line. The following examples from the *Personality Rating Scale*² of the American Council on Education for rating students illustrate the graphic rating technique:

*No Opportunity to
Observe*

A. How are you and others affected by his appearance and manner?



B. Does he need constant prodding or does he go ahead with his work without being told?



² *Personality Rating Scale*, Committee on Personality Measurement, American Council on Education, Washington, D. C., 1929.

C. Does he get others to do what he wishes?

Probably unable to lead his fellows	Lets others take lead	Sometimes leads in minor affairs	Sometimes leads in important affairs	Displays marked ability to lead his fellows; makes things go
--	--------------------------------	---	---	---

2. **The Questionnaire.** As is true of the rating scale, the questionnaire is a technique commonly used in obtaining quantitative estimates of traits not adaptable to direct testing. The chief difference is that the questionnaire represents a systematic report of attitudes, beliefs, reactions, interests, etc.; whereas the rating scale represents estimates or judgments about the possession of a trait or characteristic. The two sometimes overlap, there being some personality "blanks" which are hard to classify.

Personality questionnaires have been arranged in various forms. They sometimes consist of a list of questions to which direct answers are to be given. More often they consist of items or statements to be marked "Yes" or "No"; "Like," "Dislike," or "Indifferent to"; or to be marked in some similar fashion to indicate the subject's reaction to the items. Still others require that the person answering the questionnaire check the items applying to him, describing him or expressing his beliefs. Examples from some personality questionnaires follow:

FROM: WILLOUGHBY EMOTIONAL
MATURITY SCALE ³

(The person taking the test checks a statement if it describes him, or if it describes the person whose characteristics he is rating. *S* is the subject being rated.)

1. *S* is ordinarily friendly toward members of his immediate social group, but in critical periods becomes irritable or hostile. ----
2. *S* is extremely solicitous of his immediate family associates. ----
3. *S* makes his plans with objective reference to his own death when this issue is involved, and has no emotional reaction greater than that, for instance, concerned in planning with reference to a long journey ----
4. *S* is meticulous in matters of dress; a considerable part of his income may be spent in this activity, even though strict economies are thereby necessitated elsewhere. ----

FROM: THE HUMM-WADSWORTH
TEMPERAMENT SCALE ⁴

(The person taking the questionnaire underscores "yes" or "no" to each question.)

1. Do you like to meet people and make new friends? Yes No
5. Does noise readily waken you from sleep? Yes No
14. Have you several times been unjustly punished? Yes No
28. Has more than one person called you hot-headed? Yes No
189. Are there very few people in whom you can confide your troubles? Yes No

³ Willoughby, R. R., *Emotional Maturity Scale*, Stanford University Press, Stanford, Calif., 1931. Quoted by permission.

⁴ Humm, D. G., and Wadsworth, G. W., "The Humm-Wadsworth Temperament Scale," *American Journal of Psychiatry*, Vol. 92, No. 1, July 1935. Quoted by permission.

The Measurement of Personality 331

FROM: A-S REACTION STUDY, BY
G. W. & F. H. ALLPORT ⁵

(The person answering checks one of the choices given after each question.)

1. A salesman takes manifest trouble to show you a quantity of merchandise; you are not entirely suited; do you find it difficult to say "No"?

yes, as a rule -----

sometimes -----

no -----

2. Some possession of yours is being worked upon at a repair shop. You call for it at the time appointed, but the repair man informs you that he has "only just begun to work on it." Is your customary reaction

to upbraid him -----

to express dissatisfaction mildly -----

to smother your feelings entirely -----

3. When you are served a tough steak, a piece of unripe melon, or any other inferior dish at a high-class restaurant, do you complain about it to the waiter?

occasionally -----

seldom -----

never -----

The items of a questionnaire are usually included only after considerable experimentation with items constructed to find out about the trait or traits to be measured. If items are to be assembled to measure the degree of possession of Extrovert-Introvert Tendencies, then experimentation should demonstrate a reliable difference between answers of introverts and extroverts before the

⁵ Allport, G. W. & F. H., *A Scale for Measuring Ascendancy-Submission in Personality*, Houghton Mifflin Co., Boston, 1928. Quoted by permission.

332 The Measurement of Personality

items are included. If questions are to be used to discover neurotic trends, then it should be demonstrated prior to the final use of the questions that their answers actually differentiate neurotics from normals. This validation of items or questions involves many of the difficulties of validation of personality tests in general, but without it the questionnaires could be considered of little value.

It is customary to assign numerical values to answers on a questionnaire in accordance with the demonstrated worth of the answers in indicating the trait to be measured. Thus a total questionnaire may give a score representing degree of possession of introversion, submission tendencies, neurotic trends, or emotional maturity.

3. Objective tests. By this method personality is measured by the individual's actual performance on tasks or problems set before him in the test. The method is the method of psychological testing met almost universally in measuring intelligence and academic achievement. The Social Intelligence Test described in Chapter XX and the performance tests of the Honesty and Trustworthiness Series of Hartshorne and May described in Chapter XXII are good examples. Most of the instruments for measuring personality that are in the form of objective tests present the difficulties of point (b) of our discussion of general difficulties in measuring personality.

Relative values of personality test techniques. The rating and questionnaire methods, as compared with objective tests, usually lack reliability. By these methods we seldom find exactly the same results when the same judge rates the same group at different times; and two or more judges rarely show agreement in their esti-

mates of the same individuals. However, much of the potential unreliability of these methods can be avoided by carefully constructed scales, by conscientious, trained judges or raters, and by combining ratings of several judges. Under favorable conditions, the reliability of these methods, while below that of standardized intelligence tests, is still high enough to warrant their continued use. This is especially true in the face of the impossibility of applying objective test methods to all the traits for which measurements are needed in the course of dealing with human problems.

When the factors which reduce the reliability of ratings and questionnaire estimates are considered, the validity correlations for these methods seem fairly satisfactory—indeed, as satisfactory as the validity correlations usually obtained for objective personality tests. No technique of personality measurement has yet proved thoroughly valid. The best that we can do at the present stage of testing in this field is to accept the best of the devices of all types, wherever possible supplementing the one by the other.

CHAPTER XX

The Measurement of Social Attributes

IN THIS chapter we shall consider certain instruments of measurement which have been worked out for estimating degree of ability to get along with people or to adjust oneself in a social world—a quality often spoken of as “social intelligence.”

The value of the ability to live and to work harmoniously with one's fellow-beings, or social intelligence, hardly needs emphasizing. Lack of this quality partly accounts for many failures where abstract intellectual ability is sufficient or over-sufficient. Capacity to appreciate the feelings and attitudes of others and to carry on satisfactory relationships with them is absolutely essential in many of life's occupations. Positions that involve personal contacts require no qualification more important than the power to analyze social situations, to appreciate the forces that govern the behavior of people, to deal with subordinates, and to understand and to execute the policies of superiors.

The man is conspicuously lacking in these qualities who builds up a friendship and, as host, offends his friend; or who heads his college class in scholastic honors only to fail in the practical world; or who laboriously trains an apprentice and, because of a thoughtless outburst of temper, loses him at the beginning of his period of real usefulness.

One of the measurements that we shall discuss has been

developed for indicating social intelligence at a relatively high level of general ability; the other has been developed for measuring it over a wider range.

I. The Moss-Hunt-Omwake Social Intelligence Test

1. **The factors tested.** The separate parts of this test are based on the elements entering into the ability to get along with other people. Parts included in the various forms of the test have covered seven factors of importance in social intelligence. A discussion of these factors will indicate the purposes of the test.

One of the most important elements in social intelligence is the *ability to remember names and faces*. The person who gets along best with others does not have to be introduced to a man many times before he remembers that he has met him before. If he is a salesman, he is able to address the prospective customer by name if he calls at the place of business a second time. The pleasurable reaction in one who is recognized and called by name is realized by the politician, who attempts to call every doubtful voter by name, and by the traveling salesman, who learns the buyer's name before calling upon him.

There are two factors to be tested in this ability. One is recognition of the face that has been seen before; the other is linking the name with the face. The two ordinarily exist together, but it is not unusual to find a person who readily remembers faces, but experiences difficulty in remembering names. Results of use of the test indicate, as would be expected, that it is easier to pick out from a large number of faces those that have been seen before, than it is to link the proper names with the faces.

Another factor in social intelligence is the *ability to*

recognize from facial expressions the attitudes, states of mind, or emotions in others. The first requirement for successful dealing with others is knowledge of how the other is reacting or how he feels. One who is unable to tell by a person's expressions whether he is indifferent, mildly interested, or greatly interested, is hard put to select the most effective means of clearing up a difficult situation. One who cannot distinguish between expressions of disgust and anger may fail to choose the appropriate word at a critical moment. In short, this factor in social intelligence involves the ability to recognize from the facial expression the mental state or attitude behind it.

Closely related to the trait just described is the *ability to judge the state of mind or motive behind words.* It is clear to everyone that words of anger and words of admiration are vastly different. But very much finer distinctions between spoken expressions of mental states can be made. The finer the distinctions a person can make and the quicker he is at discerning the motive behind words, the better the words of others are as a guide to his social relations. Since words are one of the most widespread means of expressing thoughts and motives, it is not difficult to see why a person who is unusually successful in his relations with others is likely to be one who can see farthest behind the words.

Comprehension of social situations and judgment in solving the situations are of extreme importance. The motives or causes producing situations should be clear to the persons involved, and the methods of solving social problems must be available. The correct solution of a problem may be vastly different, according to whether the situation was produced in relations of friendship and pleasant motives or of enmity and unpleasant motives.

Of the various number of actions possible in a given situation there is perhaps one which would produce absolute discord, another which would leave unfriendly feelings, another which would answer satisfactorily but leave an attitude of indifference, and another which would answer just as satisfactorily and in addition build up pleasant relationships. With these possibilities in mind, the person who has the most social intelligence chooses the last.

Ability to deal with people is reflected through one's *knowledge of human nature* as obtained by observations of people under various circumstances. Everyone has had opportunities to observe human nature; to find out that certain tactics succeed while others fail. Thus observation of human behavior has been included as a factor in the Social Intelligence Test. The one who is most adept in getting along with people will be, in general, the one who is most interested in people and has most accurately observed their behavior.

Breadth of knowledge, or diversity of interests, is another important factor. Dealing with people successfully means being interested in the things in which they are interested. The more things with which a person is familiar, the more likely he is to have something in common with others and therefore be better able to get along with them. Since fields of interest are so varying, the person with a knowledge of the greatest number of fields can appeal to the greatest number of people. The person with a high degree of social intelligence is not at loss for words if the conversation happens to be outside of his own limited vocation. Diversified interests, moreover, are valuable not only in themselves but also indirectly in many of the factors mentioned before: judgment, for example, is limited by interest and knowledge.

Lastly, *sense of humor* seems related to social intelli-

gence. One devoid of such a sense lacks an important point of contact with his fellow beings.

2. Sample social intelligence test items. The various parts of the test have been prepared to measure the above-described qualities in such a fashion as to be suitable for group testing. Samples of various of the parts are presented from the revised form of the test.

1. JUDGMENT IN SOCIAL SITUATIONS

Directions. Four answers are suggested for each of the following questions. Make a check (✓) in the space in front of the answer which you consider to be most nearly correct. Do not check more than one answer. If you do, your work on that question will not be counted.

1. A man who has been a traveling salesman for fifteen years decides, under pressure from his family, that he will stay in one place and is transferred to the general office of his company. You would expect him to:

----- Like the office work because it is restful.
 ----- Become restless under office routine.
 ----- Seek a position with another firm.
 ----- Be very inefficient in his office work.

2. You wish to ask a favor of an acquaintance whom you do not know very well. The best way to ask him would be to:

----- Try to impress upon him that he is the one who will benefit.
 ----- Tell him how greatly he can benefit you if he does it.
 ----- Offer to do something for him in return.
 ----- Ask him, briefly stating your reasons.

3. Assume that you are a girl and that you meet an older woman on the street, a very slight acquaint-

ance, whose eyes show evidence of weeping. It would be best to:

- Ask her why she is sad.
- Put your arm around her in a consoling manner.
- Appear not to notice her distress.
- Appear not to see her at all.

4. Suppose you are president of a community center organized for improving community conditions. Meetings for the last three months have been poorly attended. The best way to bring more citizens to the next meeting would be to:

- Visit some of the prominent citizens and lay some of the new problems before them.
- Advertise an interesting program for an evening meeting.
- Post notices of the meeting in all public buildings.
- Send a personal notice of the meeting to all members.

5. Smith, an efficient employee, but one who thinks he "knows it all," has for several days been criticizing Jones' method of doing a particular piece of work they are doing. The boss later suggests that all the workers follow Jones' method, since it saves time. Smith will most likely:

- Ask the boss for another job.
- Keep on doing it his own way without comment.
- Follow Jones' method but continue to criticize it.
- Follow Jones' method but purposely work less efficiently.

6. Suppose you are a junior clerk in a large office. The chief of your section enters your room while you are reading a newspaper when you should be working. The best way out of the situation would be to:

- Continue reading the newspaper and show no embarrassment.
- Fold it up and return to your duties.
- Appear to be making news clippings relative to your work.
- Try to interest the boss by reading aloud an important headline.

2. RECOGNITION OF THE MENTAL STATE OF THE SPEAKER

- | | |
|-------------------|----------------|
| 1. Ambition | 10. Hypocrisy |
| 2. Admiration | 11. Indecision |
| 3. Despair | 12. Jealousy |
| 4. Determination | 13. Loneliness |
| 5. Disappointment | 14. Love |
| 6. Disgust | 15. Rage |
| 7. Envy | 16. Regret |
| 8. Fear | 17. Scorn |
| 9. Hate | 18. Suspicion |

Directions. In the parentheses before each of the following quotations write the number of the word from the list above which most accurately describes the mental state of the person making the statement. Some of the mental states in the list may not be represented at all below, and some may be represented more than once.

- () No one is able to stop me: I will do that which I intended to do or die in the attempt.
- () There is something in the way he deals that makes me want to cut the cards.
- () A glance from your eyes, a touch of your hand, and the gates of paradise swing wide for me.
- () Drink as much wine as you please but preach the benefits of water.
- () The idea of asking those Baileys! They wouldn't even know a reception from a strawberry festival.
- () I wish I had your opportunity. Things are always handed to you on a silver platter, but I never get a chance to do anything.

- () Now could I drink hot blood and do such bitter business as the day would quake to look on.

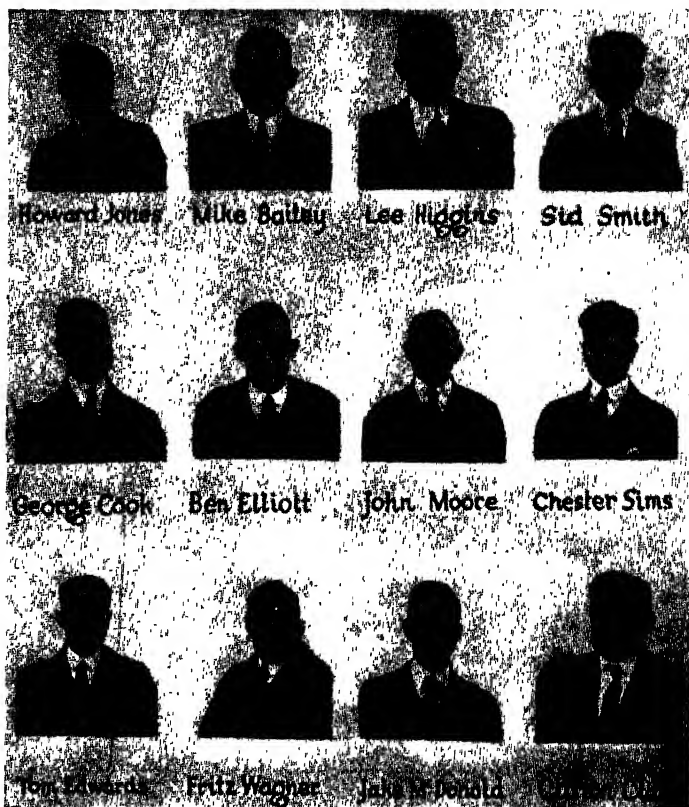
3. OBSERVATION OF HUMAN BEHAVIOR

Directions. If the statement is true, encircle the *T*; if it is false, encircle the *F*.

- T F 1. In pleasure the corners of the mouth are pulled down.
- T F 2. Pretense and sham are often inspired by the desire for social admiration.
- T F 3. Most people tend to imitate those whom they admire.
- T F 4. It is easier to remember to wind an eight-day clock than one that must be wound every day.
- T F 5. All men are created equal in mental ability.
- T F 6. We are more shocked by our errors in etiquette than by those in logic.
- T F 7. In fear there is a tendency for the eyes to become more widely opened.
- T F 8. As a rule we should place little confidence in those who appear to love us extremely on a slight acquaintance.
- T F 9. A person of strong character usually makes firm friends and bitter enemies.
- T F 10. For most people, forbidding an act increases the pleasure of doing it.
- T F 11. A mother's estimate of her child is the most reliable one.
- T F 12. Good conduct is a reliable indication of high intelligence.
- T F 13. The salesman who makes the most sales is usually the most popular with the other salesmen.
- T F 14. One of the surest methods of bringing a man to your point of view is by engaging in argument with him.
- T F 15. With the average person there is no more pleasing sound than praise of himself.

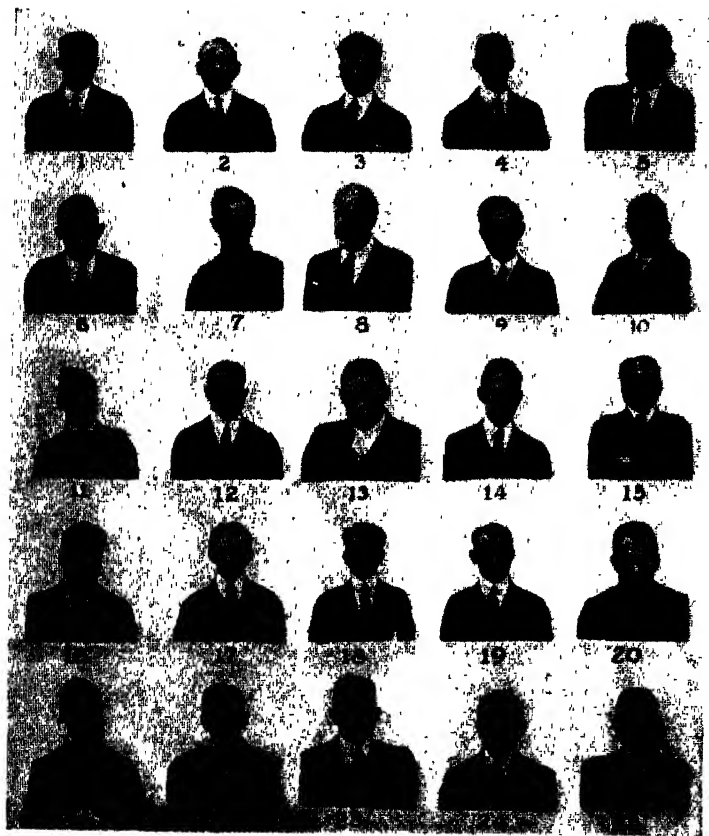
STUDY SHEET—MEMORY FOR NAMES AND FACES

Study each of the twelve faces very carefully and try to remember the name that goes with it, for later you will have to recognize these faces in a larger group and remember their names. You will have four minutes to study this sheet.



TEST—MEMORY FOR NAMES AND FACES

Included in the photographs below are the 12 individuals whose photographs and names you saw on the sheet which you studied at the beginning of the test. Below are the names of these twelve individuals. From the group below pick out the 12 persons whom you have seen before and write the number of each of these 12 in the parentheses before the name with which it goes.



- () Mike Bailey.
- () Clifton Clark.
- () George Cook.
- () Tom Edwards.
- () Ben Elliott.
- () Lee Higgins.

- () Howard Jones.
- () Jake McDonald.
- () John Moore.
- () Chester Sims.
- () Sid Smith.
- () Fritz Wagner.

4. SENSE OF HUMOR

Directions. In each of the following, pick out the one of the four suggested completions that makes the best joke, and write its number on the line at the right.

1. Cable to Scotchman whose wife had been lost at sea after a shipwreck: "Wife's body found. Attached to it a rare fish for which the British Museum offers four hundred pounds." Reply: (1) "Accept the offer. Reset the bait." (2) "Sell fish to highest bidder." (3) "No more fish stories." (4) "My wife always said she could catch anything." -----
2. Physician while taking case history asks, "Are you married?" Patient: (1) "Yes, but I pay the bills." (2) "That was twenty years ago." (3) "My wife chooses her own doctor." (4) "No, the reason I look this way is because I'm sick." -----
3. Lady in a lower berth of pullman train, on being annoyed by the snoring of man above, tapped on the upper wall. Answer from above: (1) "Who paid for this upper berth?" (2) "All right, close your window." (3) "I don't like your snoring either." (4) "Sorry, lady, I saw you when you came in." -----
4. The Judge: "Wife desertion, Rastus, is a terrible thing, about which I feel very strongly and for which I must punish you severely." Rastus: (1) "I ain't a deserter, I'se a refugee." (2) "I'se only takin' a rest for my health." (3) "I won't desert her again." (4) "There ain't no justice." -----
5. "Johnny, if you eat more cake you'll burst." (1) "Why I've eaten this much before." (2) "No, I have a tough stomach." (3) "Then I'll be able to take still more." (4) "Well, pass me some and get out of the way." -----

6. She: "Since you pride yourself on being able to judge a woman's character by her clothes, what would be your verdict of my sister over there?" (1) "Short-sighted." (2) "Acquitted." (3) "Insufficient evidence." (4) "She has an alibi." -----
7. "Eliza," said a friend of the family to the old colored washer woman, "Have you seen Miss Edith's fiance?" (1) "Oh, yes ma'am, he sho' do look like Gary Cooper, don't he?" (2) "No ma'am, he don't come here 'ceptin' Wednesdays and Saturdays." (3) "No ma'am, it ain't been in de wash yet." (4) "No, I always borrows money from my friends." -----

3. **Analyses of the social intelligence test.** (a) *Its reliability.* Results obtained in using the Social Intelligence Test indicate that its reliability is sufficiently high. One hundred college sophomores took the test twice, four months apart, the test being administered by two different persons. Scores on the two different trials correlate 0.89. For 129 college students in another university, when scores on odd questions are correlated with scores on even questions and the Spearman prophecy formula (Brown's) is used for predicting reliability of the whole test, the reliability is 0.88. Scores on two different forms of the test correlate 0.85.

(b) *Its Validity.* To know the extent to which the test actually measures ability to deal in human relationships, it is necessary to have some measure of social intelligence with which to compare scores. As would be expected, it has been difficult to find such a measure sufficiently objective and reliable to use in correlating scores. Since very few quantitative studies of this ability have been made, largely because of lack of means

to measure objectively "adeptness in handling people," adequate measuring devices for establishing a criterion have not been easily available.

One means of obtaining a definite measure of complex traits for which it is impossible to obtain test measures is through personal ratings by competent persons who know the individuals to be rated. In a large sales company 98 employees who took the Social Intelligence Test were rated by a superior executive who had good opportunity to know their ability to deal with people. The ratings on a seven-point scale correlated 0.61 with Social Intelligence Test scores. Of those making above-average scores on the test, 75 per cent, as rated by the judge, were above average in ability to get along with people.

Study has also been made of the relationship between test scores and ratings of students, student ratings representing a combination of teacher ratings and sorority or fraternity ratings. The study gave positive results, but somewhat lower than for the industrial rating study; correlations averaged about 0.40.

Studies of the Social Intelligence Test have also considered, as a criterion of the ability to be measured, the extent of participation by students in extra-curricular activities. In one study of 262 freshmen taking full-time college work, the activities listed for study are those in which participation is entirely voluntary and aside from class work, including the following: athletics, dramatics, literary society, class offices, student publications, debate, glee club, fraternities or sororities, and participation in social functions. The assumption has been that extra-curricular activities are an indication of the sociability of the student or the skill with which he deals with his fellow students. This has been considered a safe

assumption, since extra-curricular activities depend directly upon association and dealing with fellow students. The student who is unable to adjust himself to the actions of others, or who does not like to engage in pursuits where he must manage relations with others, is an infrequent participant in campus activities.

TABLE XXIV

RELATIONSHIP OF SOCIAL INTELLIGENCE TEST SCORES TO
NUMBER OF EXTRA-CURRICULAR ACTIVITIES ¹

<i>Number of Activities</i>	<i>Lower Quarter</i>	<i>Upper Quarter</i>	
	<i>Point</i>	<i>Median</i>	<i>Point</i>
Four or more	105	116	125
Three	97	112	121
Two	101	110	124
One	92	105	117
None	83	99	113

Of the 262 freshman students, 90 engaged in no extra-curricular activities, 59 in one, 62 in two, 31 in three, and 20 in four or more. Table XXIV gives the scores on the Social Intelligence Test for the different groups.

On another occasion, a group of upper-class students was studied. Activity scores were calculated for the students by assigning numerical values to various types of participation as indicated by yearbook and school records, and by answers to questionnaires filled in by the students. Activity scores ranged from 0 to 50. The total scores on the Social Intelligence Test show the differentiations indicated in Table XXV.

Such studies have been interpreted as indicating that the test is measuring something of significance in determining participation in campus organizations and activities. Equal results are not obtained with other meas-

¹ Hunt, T., "The Measurement of Social Intelligence," *The Journal of Applied Psychology*, Vol. XII, No. 3, June 1928.

ures of student ability, as abstract intelligence tests, grades, and scholastic honors.

TABLE XXV
RELATIONSHIP OF SOCIAL INTELLIGENCE TEST
SCORES TO ACTIVITY SCORES ²

<i>Activity Score</i>	<i>Social Intelligence Test Average</i>
30 or over	128
20 to 29	125
15 to 19	117
10 to 14	118
5 to 9	118
1 to 4	114
0	113

4. Relationships of social intelligence as measured by the test. (a) *To other types of intelligence.* Studies of the relationship between social intelligence and abstract intelligence and concrete intelligence may throw some light upon the distinctness with which we can speak of the three types of intelligence included in Thorndike's three-fold division. Such studies may also throw some light upon the extent to which we can measure social intelligence as distinguished from the other two types, particularly as distinguished from abstract intelligence. It has frequently been suggested by critics of the Social Intelligence Test which we have discussed in this chapter that the test measures too large an element of abstract intelligence, probably mainly through its verbal medium of testing. The correlations between the Social Intelligence Test and abstract intelligence tests may help to answer this criticism. The literature contains reports on eight studies of the relationship between the Social Intelligence Test and abstract intelligence tests, all these

² Moss, F. A., "Preliminary Report of a Study of Social Intelligence and Executive Ability," *Public Personnel Studies*, Vol. IX, No. 1, 1931.

studies being done on college students. These are summarized in Table XXVI. The average of these correlations is .49.

TABLE XXVI
CORRELATIONS BETWEEN SOCIAL AND ABSTRACT
INTELLIGENCE

<i>Author</i>	<i>Tests Correlated</i>	<i>Correlation</i>	<i>No. of Cases</i>
Broom	Thorndike Test w. Social Test	.56	258
Hunt	Mental Alertness Test w. Social Test	.54	243
Hunt	O'Rourke Test w. Social Test	.57	102
Hunt	McCall Multi-Mental Test w. Social Test	.25	130
Pintner and Upshall. . .	Group Verbal Test w. Social test	.68	33
Garrett and Kellogg . . .	Thorndike Test w. Social Test	.42	118
Strang	Group Verbal Test w. Social Test	.44	311

Two studies of relationship between Social Intelligence Test scores and mechanical intelligence scores have been reported. One of these studies utilizes the O'Rourke Mechanical Aptitude Test, the correlation being .22; the other utilizes the MacQuarrie Mechanical Ingenuity Test, and the correlation is .11.

In order to draw any conclusions about the relationships between the three types of intelligence, we need also to know something of the relationship between abstract intelligence and mechanical intelligence, and something of the relationship between different tests designed to measure the same one of the three divisions. The average of numerous correlations reported between abstract and concrete intelligence, as measured by the available tests, is about .50. The correlations between various abstract intelligence tests average about .75 to .80. Rela-

tively few studies have been made correlating the various mechanical or concrete tests, and since only one or two attempts have been made to measure social intelligence, practically no studies are available correlating social intelligence as measured by several different tests. If we can take reliability studies of the social intelligence tests as indicative of their correlations with themselves, we can say that these correlations also would be around .75 to .80.

From the above studies, we may conclude that the three types of intelligence as designated by Thorndike can be somewhat differentiated, although our tests for measuring the three undoubtedly measure a large element of a common factor, which common factor seems to be in the main abstract intelligence. In other words, the great majority of tests of the pencil-and-paper type measure an element of abstract intelligence in addition to any other separate element which they may be designed to measure. It would seem, however, that the concrete intelligence tests and the one social intelligence test which we have discussed measure something sufficiently different from abstract intelligence so that they are of some value in our studies of abilities.

(b) *To sex.* Studies of the Social Intelligence Test have shown in average scores of women a small but reliable difference from scores of men of equal age and educational status. The parts of the Social Intelligence Test in which women show the greatest superiority are those parts measuring Judgment in Social Situations and Observation of Human Behavior.

(c) *To occupation.* The Social Intelligence Test has been given to a considerable number of occupational groups, and norms have been published for some of these groups. The median score attainments of several are given in Table XXVII. Studies of the relationship be-

tween adequate measures of social intelligence and of occupational success should be very valuable in connection with the selection of workers for those positions in which the ability to get along with others is of supreme importance.

TABLE XXVII
MEDIAN SOCIAL INTELLIGENCE SCORES FOR
VARIOUS OCCUPATION GROUPS³

<i>Occupation</i>	<i>Number of Cases</i>	<i>Median Score</i>
Administrative and executive employees . . .	100	117
Teachers	250	112
High-grade secretarial employees	50	111
Salesmen	25	107
Engineering employees—draftsmen, electricians, etc.	45	105
Clerical and stenographic employees	200	95
Other lower-grade office workers and helpers	300	84
Sales clerks in department store	35	81
Nurses	75	78
Lower-grade industrial workers	150	65

II. The Vineland Social Maturity Scale

Edgar A. Doll, Director of Research at the Training School for Feeble-minded at Vineland, New Jersey, has recently been working on a scale for measuring social competence. He believes that social competence is sufficiently differentiated from intelligence to make its measurement important in his work with lower-grade individuals. Study and further use of the scale will undoubtedly prove its value for application to many problems among normal and superior individuals.

We quote from Doll⁴ his statement of the problem of measurement of social competence:

³ Hunt, T., *op. cit.*

⁴ Doll, Edgar A., "The Measurement of Social Competence," *Proceedings of the Fifty-ninth Annual Session of the American Association on Mental Deficiency*, Chicago, Illinois, April 25-27, 1935.

The attempt to measure social competence immediately encounters critical difficulties. Shall we identify social competence with social success as reflected in fame or fortune? Shall we define social competence as the ability to manage oneself and one's own affairs independently, or, should we also require some contribution to the welfare of others? Social competence is affected by intelligence, by personality, by emotionality, by conduct, skill, opportunity, training, experience, and so on. But social competence is not to be measured by any of these traits alone. Social competence is measured rather by the effective social uses to which these traits are put. The ultimate significance of any human trait is its actual capitalization for social purposes. We may even go so far as to say that no behavior is important except as it is socially significant.

Social competence is also influenced by physical health and well-being, by disabling diseases, by sensory handicaps, invalidism, crippling, or senescence. Environmental influences, too, play a role; for example, the restraints of family life, the solicitude of elders, the strength of motivation, incentive, and habit, the stimulus of ideals, the customs of time and place, the advantages or limitations of cultural status, and so on. In particular, we are obliged to distinguish between sociality and sociability, meaning by the former social competence and by the latter social affability.

In constructing the Vineland Social Maturity Scale we have reckoned with these and other difficulties. The extent to which we may have succeeded in resolving these complicating influences will be reflected in the practical success of this instrument as further investigation proceeds. We have made an attempt to measure social competence as a composite capitalization of the sum total of individual capabilities which is reflected in progressive stages of personal independence or freedom from the assistance and supervision of others.

1. The Vineland scale. In its present form, the Vineland Social Maturity Scale consists of 117 items, arranged

in order of their development in normal persons from birth to adult life. Each item is designed to indicate social competence in terms of the degree to which the individual is capable of looking after himself, getting along with others, and contributing to the welfare of others. The items of the scale include such general activities as self-help, locomotion, occupation, communication, self-direction, and socialization. The manual of directions defines, explains, and illustrates each item. The scale is meant to be not a rating scale but an objective schedule of developmental behavior capable of being standardized in definite quantitative terms. The information for making a measurement is to be obtained from observation and by questioning the subject and others familiar with the subject's behavior. The method of standardization being employed by Doll is the Binet method. Results are expressed as Social Age and, by dividing by chronological age, as Social Quotient. Sample items from the middle range of the scale are quoted to show the general nature of the material.

Bathes self-assisted
 Uses table knife for cutting
 Uses pencil or crayon for drawing
 Uses skates, sled, wagon
 Goes about neighborhood unattended
 Cares for self at table
 Prints simple words
 Combs or brushes hair
 Plays competitive exercise games
 Does routine household tasks
 Goes to bed unassisted
 Plays simple table games
 Goes to school unattended
 Disavows literal Santa Claus
 Uses pencil for writing

2. Studies of the scale. The studies of the scale reported by Doll up to the present are all of a preliminary nature. He reports reliability of the scale based upon a second application as applied to 30 subjects. Fifteen of these subjects, of average social age 8.1, were reëxamined by the same examiner, with a different informant the second time. The median difference between the two examinations was $0 \pm .6$. The other fifteen subjects were reëxamined by a second examiner, who used the same informant that was used by the first examiner. The results showed a median difference of $+.2 \pm .4$. These studies on a few subjects would seem to indicate satisfactory reliability for the measuring device. Validity is less satisfactorily indicated in the preliminary reports, but the lack of a generally good criterion for validating the scale would leave us to await further study before judging the scale's validity. The preliminary results indicate a relationship between Social Age and Mental Age, and between Social Age and Cultural Status, of the subjects.

3. Uses of the social maturity scale. Doll suggests for his scale the following uses in relation to institutional administration. Extension of the uses to other problems can be made likewise. Doll's suggestions follow:

1. The Scale affords a measure of social competence which is useful in establishing the fundamental criterion of mental deficiency and, consequently, for discriminating between the feeble-minded and the intellectually subnormal.
2. The Scale affords a measure of progressive social growth, arrest, or deterioration, and, *ipso facto*, affords a schedule by means of which an adequate developmental history may be taken to distinguish between arrested mental development and deteriorated mental states.

3. The Scale affords an objective measure of relative social competence within a feeble-minded group, and suggests the extent to which special handicaps need or need not be taken into consideration.
4. The Scale affords a means of measuring improvement in social ability as brought about by treatment or training, and indicates the direction as well as the amount of such improvement.
5. The Scale meets a basic need in mental diagnosis which is of critical importance in borderline cases and provides a standard measure that may be employed for legal purposes, since the legal definitions of idiot, imbecile, and moron have always been expressed in social terms.
6. The Scale is an aid in the classification of institutional inmates in respect to housing groups, training groups, and work groups, and as a general criterion in the study of social adjustment.
7. The Scale is particularly useful in considering patients for discharge.
8. The Scale affords a means of discovering unsuspected social aptitude of those patients whose unobtrusive personality and social withdrawal tendencies render them inconspicuous and lead to underestimation of their real ability.
9. The Scale affords a useful device in mental hygiene consultation where social capability is likely to be a critical factor in relation to social adjustment.
10. The Scale affords a practical measure of insight that should be useful with abnormal patients.
11. The Scale affords a guide for child training and parent education, suggesting the direction in which training may be advisable and the limits within which success is likely to be achieved in relation to ability.
12. The Scale affords a definite advantage over personal methods of estimating social competence.

13. Finally, the Scale provides a basic criterion for the investigation of many different problems in social science which have heretofore been complicated by lack of a suitable criterion.

CHAPTER XXI

The Measurement of Emotions by Verbal Tests

IT IS rather difficult to give an exact definition of "emotions" as they are measured by the types of tests to be considered in this chapter. Generally the tests have to do with our tendencies to emotional reactions. The results or the scores on the test indicate that we react with emotion to more or fewer things than the average person of our age and sex does; or they indicate that in us emotional reactions are provoked by those things which ordinarily provoke emotions, or by peculiar things which ordinarily do not call out emotions; or they may indicate that we experience more intense or peculiar kinds of emotional reactions to those things producing emotions.

We may think of emotions as being tested in two general ways—by a direct method and by an indirect method. The *direct method* consists of putting the subject to be tested in a situation which is calculated to arouse emotional responses, and then measuring his reaction under these conditions. Such measurements have been carried out numerous times in the psychological laboratory, but the procedures and means of measurement are often cumbersome, time-consuming, and unstandardized, so that direct measurements of emotions have not reached a stage of general application. From the standpoint of validity they probably have much to recommend them over the

indirect methods. Chapter XXVI is concerned with some of these measurements.

The *indirect method* is usually some verbal type. Words are used as stimuli, and the individual being measured is asked to respond by words. The method may investigate the reaction which the individual reports that he experiences when he thinks of the word. Or an imaginary situation may be set up and the individual may be asked to report what his reaction would be in the imagined case. The indirect methods usually involve a certain amount of introspection and retrospection. The examiner does not observe the subject's reactions directly; he can observe and examine only the reports of reactions.

I. Adjustment Questionnaires and Tests

Several scales or tests have been worked out for measuring emotional adjustments. These usually aim at discovering abnormalities of reaction, idiosyncrasies or peculiarities of emotions, or neurotic traits. The items of many of them are simply scored for normality or abnormality of response. Practically all of them have their chief value and usefulness in the discovery of abnormalities or neurotic traits, and are of only limited, if of any, usefulness within the normal range of behavior. Examples of these are the following:

1. **Personal Data Sheet.**¹ This is a questionnaire constructed by Woodworth for discovering an individual's psychoneurotic tendencies. It consists of 116 questions to be answered by *Yes* or *No*, such as: "Does it make you uneasy to go into a tunnel or subway?" "Do your interests change quickly?" The score is determined by the num-

¹ Published by the C. H. Stoelting Company, Chicago, Ill.

ber of neurotic or unfavorable answers, the unfavorable answer sometimes being *Yes*, sometimes *No*. The scale is suitable for adults, being originally designed for use with soldiers. It is one of the earliest of the adjustment scales, and at present there are available a number of revisions based upon Woodworth's original Data Sheet. Ellen Mathews published in 1923 a revision suitable for children eight years of age and over. S. D. House published in 1927 a revision for adults, which he called a Mental Hygiene Inventory.

2. Personality Schedule.² This is a questionnaire very similar in form and scoring to the Woodworth scale. It contains a somewhat larger number of questions, 223 in all. It is designated for discovering personal and social maladjustments, particularly in college students. The norms give score ranges for various degrees of maladjustment from the normal to the definitely maladjusted.

3. Colgate Mental Hygiene Test.³ The test is designed to detect abnormal emotional trends in adults. It has been most used with college students. The test is in the graphic scale form; answers to the questions are to be indicated by checks along a line beneath which are terms or phrases descriptive of the various possible reactions or answers. There are three parts to the scale: Part I, containing questions dealing with psychasthenia; Part II, containing questions dealing with schizophrenia (split personality); and Part III, containing questions dealing with neurasthenia. Checks in the "neurotic" end or section of the scale are scored. The total score represents the extent of abnormal trends.

² By Thurstone, L. L. and T. G. Published by University of Chicago Press, Chicago, Ill.

³ By Laird, Donald. Published by Hamilton Republic, Hamilton, N. Y.

4. Personality Inventory.⁴ This test contains 125 questions to be answered by *Yes*, *No*, or *?*. The first 15 questions are reproduced here as samples:

1. Yes No ? Does it make you uncomfortable to be "different" or unconventional?
2. Yes No ? Do you day-dream frequently?
3. Yes No ? Do you usually work things out for yourself rather than get someone to show you?
4. Yes No ? Have you ever crossed the street to avoid meeting some person?
5. Yes No ? Can you stand criticism without feeling hurt?
6. Yes No ? Do you ever give money to beggars?
7. Yes No ? Do you prefer to associate with people who are younger than yourself?
8. Yes No ? Do you often feel just miserable?
9. Yes No ? Do you dislike finding your way about in strange places?
10. Yes No ? Are you easily discouraged when the opinions of others differ from your own?
11. Yes No ? Do you try to get your own way even if you have to fight for it?
12. Yes No ? Do you blush very often?
13. Yes No ? Do athletics interest you more than intellectual affairs?
14. Yes No ? Do you consider yourself a rather nervous person?
15. Yes No ? Do you usually object when a person steps in front of you in a line of people?

This scale is one of the most recent of the adjustment type of tests, and makes use of considerable material demonstrated to be of value in the older scales. The

⁴ By Bernreuter, R. G. Published by Stanford University Press, Stanford, Calif. Sample questions quoted by permission.

scoring of the questions is somewhat more refined than that for the older scales. To each question weights have been assigned varying from plus 7 to minus 7 in accordance with the diagnostic value of the question. The inventory can be scored by four separate scales or scoring devices. By Scale 1 the final score indicates neurotic tendencies; by Scale 2, self-sufficiency; by Scale 3, introversion-extroversion; and by Scale 4, dominance-submission. The correlation between Scale 1 and 3 scores is very high, however, so that little is gained by using both of these scorings, and the author recommends for general use that only three scorings be done. Standards are available on the test for high school students, college students, and adults.

Upon what basis should we judge the usefulness of such measurements as those given by the adjustment scales? The various scales have usually been found to be reliable, giving approximately the same results when applied more than once at short intervals to the same group. Most of the reliability coefficients reported by the authors are close to .90. The validity of a scale of the type we have just discussed depends primarily upon the working out of the scoring device. The classification of answers for scoring as representative of adjustment or unadjustment, normality or abnormality, must be based upon studies conducted among trial groups which have been well selected for normality and abnormality. Herein lies a difficulty that is hard to overcome entirely in the construction of such scales. For most of the scales studies have been reported in which the relationship of total scores as finally rated to neurotic trends or unadjustment trends is analyzed. Mathews⁵ reports correlations of .52 and .66

⁵ Mathews, E., "A Study of Emotional Stability in Children," *Journal of Delinquency*, Vol. 8, 1923, pp. 1-40.

between scores on the Personal Data Sheet and composite ratings on "nervous instability" for groups of girls in a protectory. The ratings were made by four competent judges. Cady⁶ reports correlations averaging .40 between Personal Data Sheet ratings and estimates of incorrigibility made on a group of 150 boys in institutions for delinquents.

The various scales for measuring emotional adjustment show fairly high agreement among themselves. While such an agreement is not proof of validity, it portrays a much more hopeful situation than would be true if the various scales designed for measuring the same traits showed marked disagreement. Bernreuter reports a correlation of .94 between his Inventory scored for neurotic traits and Thurstone's Personality Schedule. When scored for introversion, his Inventory shows a correlation of over .75 with other tests for introversion.

These various studies certainly suggest value in such measurements. The limitations of the scales seem to be primarily in their lack of fine discrimination, particularly at levels other than those of the definitely maladjusted. However, we might expect a more extended use of the scales to demonstrate their practical value in vocational guidance, clinical work, and other types of personality study.

II. Extroversion-Introversion Scales

Because of the importance of tendencies to extroversion or introversion in the affairs of life, a number of measuring devices have been worked out for indicating these tendencies. We have already noted that Bernreuter's scale

⁶ Cady, V. M., "The Estimation of Juvenile Incorrigibility," *Journal of Delinquency Monographs*, Vol. 2, 1923.

can be scored for introversion. Other tests along this line include Laird's Personal Inventory, C2 and C3; Allport's Ascendancy-Submission Scale; Neyman's and Kohlstedt's Personal Traits Rating Scale.

Extroverts are individuals who express their emotions in action and association with others. They are characterized by sociability, interest in others, lack of worrying or moodiness, lack of self-consciousness, self-sufficiency, and fluency in talking. Introverts, on the other hand, are those whose emotions are expressed largely within themselves. They are characterized, in contrast to the extroverts, by lack of sociability, self-consciousness, enjoyment of more or less solitary types of recreation, day-dreaming, easily hurt feelings, more tendency to worries, and fewer more intimate friends. Tendencies in one or the other of these directions of reaction are noticeable within "normal" groups of individuals. We can guess that such tendencies may be closely related to the type of job or vocation in which one can best succeed, and may have a great influence on one's social adjustments in his group of friends and associates. Extremes in either direction of introversion or extroversion become abnormal, and there are definite psychotic or insane conditions which typify the extremes. Dementia praecox and involution melancholia are examples of extreme introversion among the insane; mania, of extreme extroversion.

Laird, working at the Colgate University Psychological Laboratory, has done a considerable amount of work on the measurement of extroversion-introversion, and we may discuss his work as typical of measurement in this field. He has developed two scales for measurement, one for rating to be done by another individual, and one for self-rating. Both scales use the graphic rating technique. The first page of one of these scales is reproduced on this

page.⁷ For each question or trait rated, one end of the rating line indicates an introvert tendency and the other end indicates an extrovert tendency. The score is the

Personal Inventory <i>Revision</i> C 2					<table border="1" style="width: 100%; border-collapse: collapse;"> <tr><td style="padding: 2px;">Act Intro</td></tr> <tr><td style="padding: 2px;">Think Intro</td></tr> <tr><td style="padding: 2px;">Total Intro</td></tr> </table>	Act Intro	Think Intro	Total Intro
Act Intro								
Think Intro								
Total Intro								
<p>.....</p> <p>Date</p> <p>Place</p> <p><i>Directions:</i> Describe yourself by answering the questions asked in the large type. Do this by making a check mark through one of the half-inch lines. After reading over the phrases in small type after each question, think back over your life for the <i>past few months</i> and determine where the check mark should be made to describe yourself accurately. Describe your <i>average behavior and thoughts</i>.</p> <p>There is no time limit to this. Read each line entirely before making a check mark. Describe your <i>average thoughts and behavior for the past few months only</i>. Check only one of the half-inch sections in answer to each question.</p>								
1. How steadily have you worked at the ordinary tasks of the day?	continuously until convinced	steadily only at oc- casional times	shifted only with mood intensity	shifted from one to another	spurts of work			
2. How have possible mis- fortunes entered into your thinking?	worried about them for some time	worried but soon forgot	usually thought what might happen	few apprehensions	impenetrable			
3. How easily have your feelings been hurt by re- marks or actions referring to you?	easily and often resented	feelings easily hurt	usually noticed other's intentions	affected in rare instances	gave them no attention			
4. How have you considered the feelings of others?	told truth regard- less of feelings	sometimes forgotful of feelings	considered only of some people's	frank, yet tactful	generally careful not to hurt them			
5. How have you acted and felt at social affairs?	usually kept in background	trying at a few at times	let others entertain, talk, etc.	liked to be entertaining	usually lead			
6. How well have you re- membered most of the errands and details of your daily routine?	continually for- getting	occasionally forget- ful	forgot minor ones	usually remember all	rarely forget			
7. In social conversation how have you been?	talkative	on easy talker	talked when necessary	preferred listening	refrained from talking			
8. How have you decided upon matters of daily con- duct?	determined almost everything	searched for good reason in most instances	determined only important things	followed first impulse in many instances	usually followed first impulse			
9. How have you generally been about making loans?	gave to anyone who asked	lended only to close persons			rarely loaned			

Published by Hamilton Republic, Hamilton, N. Y.
 Copyright 1925 by Donald A. Laird, Colgate University Psychological Laboratory
 Publication or quotation prohibited to prevent invalidation

First Page of Laird's Personal Inventory, C 2

number of checks occurring on the introvert end of the line, so that total score is representative of degree of in-

⁷ From *Personal Inventory, C 2*. Reproduced by permission of the Hamilton Republic, Hamilton, N. Y.

troversion. Laird selected the questions for his scale from a large number of questions which he studied for their differentiating value in separating trial groups of definitely known extroverts from groups of definitely known introverts.

Studies in colleges and industries utilizing the Laird test have shown fairly high correlations with other tests of introversion; moderately high negative correlations between the test scored for introversion and that scored for extroversion; higher average introversion scores for women than for men; very little relationship between introversion and general intelligence; higher average scholastic performance in school by introverts than by extroverts; slightly increasing introvert scores with age; no constant racial differences on the test; tendencies toward greater extroversion in groups of foremen and executives; and tendencies toward greater introversion in groups of office workers, clerks, stenographers, accountants, and research workers.

III. An Annoyance Test ⁸

Cason has constructed a very interesting test in the field of emotional testing. His study and test concern the situations and stimuli which produce certain feelings of unpleasantness or annoyance, and the relations between the stimuli and the responses. In his discussion of the reasons for development of such a test Cason ⁹ has emphasized the marked individual differences among people in the kinds of feelings they experience under the same external

⁸ Published by the C. H. Stoelting Company, Chicago, Ill. Quotations from the test are made by permission.

⁹ Cason, H., "An Annoyance Test and Some Research Problems," *Journal of Abnormal and Social Psychology*, Vol. XXV, No. 2, July-Sept., 1930, pp. 224-236.

conditions; and the relationship of these individual differences to individual happiness, mental and emotional maladjustments and abnormalities, success and satisfaction in one's vocation.

The test itself consists of a list of 217 items (annoyances) to be rated according to the following scale:

- 3—Extremely annoying
- 2—Moderately annoying
- 1—Slightly annoying
- 0—Not annoying
- x—Have not been in the situation

The first fifteen items are quoted as samples:

- () 1. A person behaving in an affected manner.
- () 2. A person with a gushing manner.
- () 3. A person losing his temper.
- () 4. A person habitually arguing.
- () 5. A person in an automobile I am driving telling me how to drive.
- () 6. To see a person who is driving an automobile take unnecessary chances.
- () 7. To see a boisterous person attracting attention to himself in public.
- () 8. To hear a person talking in an unnecessarily loud voice.
- () 9. A person continually trying to borrow some of my things.
- () 10. To hear a person chewing gum loudly.
- () 11. A child not obeying his father or mother.
- () 12. A mother continually correcting her child in public.
- () 13. To see a person's nose running.
- () 14. To see a person blow his nose without using a handkerchief.
- () 15. A person not covering his mouth when he coughs or sneezes.

Cason's original material was collected from items which large numbers of people listed as things that annoyed them. Somewhat over 500 statements of annoyances were derived from this original material. These were printed and submitted for rating to several hundred subjects of all ages between 10 and 90 years. The final selection of items was based upon the criteria of (1) frequency of the annoyance in everyday life, (2) age distribution of the people who were annoyed by the thing or activity, (3) objectivity, (4) universality, (5) permanence, and (6) psychological and social significance. Initial results reported for the test show slightly higher average annoyance scores for females than for males in all age groups; and annoyance scores slightly increasing with age up to 40 to 60, then slightly decreasing up to 90.

IV. The Pressey X-O Test ¹⁰

An ambitious attempt to work out a test of emotions was made some years ago by Pressey. He designated his test as *a group scale for investigating the emotions*. His scale emphasizes the measurement of abnormal mental attitudes and pathological emotional conditions, and has been found of use primarily in dealing with abnormal groups of various types. His test is in four parts.

Part I is a test for discovering various types of unpleasant feeling. The test consists of lists of words, This test aims through the association method to discover instructed to cross out every word which is unpleasant, and finally to select from each list the one word which is most unpleasant to him. The words are so chosen that by an analysis of his answers an individual may be

¹⁰ Published by C. H. Stoelting Company, Chicago, Ill. Quotations from the test are made by permission.

identified as to his particular type of fears. Each line contains a word belonging to each of four classes—disgust, fear, sex, and self-feeling—in addition to one neutral word termed a “joker,” to indicate whether the individual is following the instructions. The first four lists are as follows:

1. disgust fear sex suspicion aunt.
2. roar divorce dislike sidewalk wiggle.
3. naked snicker wonder spit fight.
4. failure home rotting snake hug.

Part II is a modification of the free-association test. five words being contained in each list. The subject is pathological, abnormal, and criminological attitudes. The test consists of key words (given in capital letters) each of which is followed by five words which are to be crossed out if connected in the mind of the subject with the word in capital letters. The first four, as samples, are as follows:

1. BLOSSOM flame flower paralyzed red sew.
2. LAMP poor headache match dogs light.
3. BATH naked choke tree alone danger.
4. KING father baseball queen rights razor.

Part III is an ethical discrimination test. It is adapted from an earlier test worked out by Fernald. The test consists of lists of words in which the subject is instructed to cross out everything which he thinks is wrong, or which he thinks that a person should be blamed for. Finally, he is instructed to encircle the one thing in each list which he thinks is the worst. The first four lists are as follows:

1. begging swearing smoking flirting spitting.
2. fear hate anger jealousy suspicion.

3. dullness weakness ignorance innocence meekness.
4. careless fussy reckless silly childish.

Part IV is a test for discovering anxiety tendencies. The subject is told to cross out all the things in each list about which he has ever worried, and finally to circle the thing in each list about which he has worried the most. The first four lists are as follows:

1. injustice noise self-consciousness discouragement germs.
2. clothes conscience heart-failure poison sleep.
3. sickness enemies money blushing failure.
4. falling queerness religion dizziness boss.

As in Part I, the words in Part IV are arranged so as to make it possible to discover particular types of anxiety tendencies. Each list contains one word representing the following anxiety attitudes—paranoid, or suspicion; neurotic; self-conscious or shut-in; melancholic or self-accusatory; and hypochondriacal. In the first list, for example, the words in order represent these five types of anxiety. In the case of this test and also in the case of Part I, the particular pathological types can be recognized only in those cases where they are definite enough to color the answers throughout the test. The marking of a paranoid word in only one list would not be significant, but if the words throughout the test which are marked are consistently of the paranoid type, they might indicate considerable about the subject's emotional attitudes.

From his emotions test, Pressey derives two general scores. One of these he terms the *total emotionality score*, which is the sum of all the words crossed out in the whole test. Presumably, the person who crosses out a large number of words is more easily aroused emotion-

ally, has more definite emotional attachments, and is perhaps more unstable emotionally than the person who crosses out only a few words. The second score which Pressey derives is an *idiosyncrasy score*, which indicates the extent to which a subject's responses are those of the normal or average person. This score represents the total number of the circled words in each line which differ from the modal choices of the average normal person.

Several studies have been made utilizing the Pressey test. Reliability of the test as reported in the studies seems fairly satisfactory. McGeoch and Whitely, using college students as subjects, report reliability correlations for the four parts given 48 hours apart. All the coefficients are over .80. With longer intervals between the testings, their correlations are somewhat lower. Flemming reports a reliability correlation of .97 for the total test scored for total emotionality, and .50 scored for idiosyncrasy, for college freshmen.

As to the value of the test, the various studies which have been made leave us with no satisfactory conclusions. Our general impression is likely to be that the test measures a composite of many factors and that a total emotionality or idiosyncrasy score means very little, if anything. In fact, Pressey himself frankly stated in making his scale that he had assembled a miscellaneous group of material for experimental study. His own statement is significant:

The scores on the entire examination are the blurred result of a number of factors, and are of relatively little importance. However, it is possible, from the mass of data yielded by the examination, to combine certain items in such a way as to obtain, from the single examination, highly differential information with reference to a number of problems.¹¹

¹¹ Pressey, S. L., "A Group Scale for Investigating the Emotions," *Journal of Abnormal and Social Psychology*, 16: 55-64 (April 1921).

Several studies of the test have been made on criminal as compared with non-criminal individuals. The various studies indicate that total scores on the test do not distinguish criminal from non-criminal groups. However, criminals show marked deviations from the normal on separate elements of the test, and considerable information of value may be obtained by a detailed analysis of the responses. Chambers has analyzed the test for its possibilities of distinguishing the emotionally mature from the emotionally immature, and has found that some of the items constitute very good tests of emotional maturity. This same investigator also studied the test for its possibilities in the differentiation of college students as to achievement. He found certain items in the test which furnished reliable means of differentiating between those who were good students and those who were poor students. For example, he found in Part IV (things worried about) that the words "books," "self-conscious," "accidents," "rivals," "parties," were marked significantly more often by good students, and the words, "work," "failure," "police," "wrecks," and "dreams," were marked significantly more often by poor students.

From the evidence that has been collected about the Pressey Test, we may conclude that it contains considerable material which may be of value in diagnosing various emotional and conduct trends. However, the specific elements of the test which are of value in various problems can be discovered only by a detailed study and analysis of the test on the groups in question. Total scores on the test seem to represent such a complex combination of various sorts of reactions and tendencies that they indicate very little of practical value.

CHAPTER XXII

The Measurement of Character and the Moral Sense

FOR purposes of discussion in this chapter we shall define character and morality in terms of the extent to which a person adheres to the established codes of good conduct and of social and moral standards. We shall not inquire to what extent such conduct depends upon innate qualities or upon acquired reaction tendencies. There is the hypothesis, adhered to by some, that individuals differ inherently in their sensitiveness to moral distinctions and in their disposition to subject themselves to codes of conduct. On the other hand, we may take the view that, aside from a minimum of intelligence as a basis for learning, all morality, character, and conduct conformities are the product of experience, environment, or training.

Reliable and valid conduct and character measurements should have a wide field of usefulness. In the field of *crime* they should give us a yardstick for measuring an important element in the makeup of the delinquent. We cannot expect society to solve the problem of crime without methods for studying the criminal. From investigations already made, it seems fairly conclusive that mental ability and achievement measures

alone are inadequate. Perhaps we can expect a valuable addition to our instruments of measurement in well worked out character tests. In respect to *insanity*, character testing may help us to recognize early signs of breakdown in behavior, to arrive at better judgments as to degree or seriousness of the disorder, or to make a better prognosis as to chances of recovery. Measurement of *vocational competence* could well include conduct and character tests. Up to the present, vocational measurement has paid most attention to measures of ability, which, to be sure, are of prime importance in the selection of a worker. But the worker must not only possess ability to do the work at hand; he must also have certain character qualities of trustworthiness, regularity, and the like, to be of great value to his organization. Another field in which character measurement is needed is in the *schools* in connection with character education, social development, and teaching of good citizenship. Our modern schools are not interested alone in academic attainments. They are interested in certain other educational outcomes reflected in better character and greater social values. These outcomes should not be neglected in the school measurement program.

Two general methods have been used in measuring character and morality, and two different types of tests have been constructed. The one type of test is a *behavior* test. It involves actually placing the subject to be tested in a situation in which he has a choice of behavior, in which he may react with good action or bad action, as with honesty or dishonesty. The individual's response in the situation is the basis for grading him. Such a method of measurement obviously has the disadvantages of cumbersomeness in testing and often of artificiality in the test situations. The second method

consists in the devising of tests of *conduct knowledge and judgment*.

The all-important question to be answered before we use such instruments of measurement concerns the relationship between knowledge and conduct. Two authorities on psychological testing have stated the problem as follows:

Very little experience of life is necessary to make one aware that a person may give lip service to a moral principle, but may repeatedly violate it in his conduct. One's ability to state a principle, or to pass a correct judgment as to what should be done in a specific situation, therefore, is no guaranty that he would act in accordance with the right principle. It is difficult to set up a moral test, however, in which the individual shall be put in a situation which demands a moral choice. Because a test which demands verbal judgment is so much easier to administer and to devise, we may inquire further whether such a test may not have some value. Upon further consideration it appears that while a test which demands verbal response does not guarantee what the conduct of an individual shall be, it does give some information on the negative side of the case. If a person shows that he does not recognize a moral principle, we may be reasonably certain that he will not act in accordance with it. A merely verbal test may, after all, then, be of some value.¹

It seems only natural that men should have turned to tests of knowledge and reasoning for the diagnosis of conduct. Even the popular belief that action follows knowledge, that we reason out beforehand the course of action we are to pursue, gives warrant enough to investigators to experiment with tests of this type. This belief permeates our institutions, our educational theories and practices, and indeed all our relations with our fellows. Courts of criminal law decide responsibility for an act

¹ Freeman, Frank N., *Mental Tests*, Houghton Mifflin Co., Boston, 1926, p. 214.

on the basis of the defendant's ability to discriminate between right and wrong.²

The first set of tests of the *behavior* type was devised and used by Voelker.³ There were two series of ten tests each. His studies of the test included validity studies based upon correlations between test results and teacher estimates of trustworthiness of students; and studies of effects of several weeks' instruction in trustworthiness upon scores. His early results were encouraging if not conclusive. The pioneer work in character tests of the knowledge and judgment type is represented in Fernald's *Ethical Discrimination Test*,⁴ in which individuals were required to rank misdeeds in order of their gravity.

The best and most extensive investigations of character measurement published to date are those of the Character Education Inquiry conducted by Hartshorne and May and their collaborators. We shall briefly examine their tests as samples in this field. For a complete consideration of the tests the reader is referred to the original reports.⁵

I. The Tests of Conduct Knowledge and Judgment

The Character Education Inquiry made use of twelve tests of the knowledge and judgment type. These are as indicated on the following page.

² Symonds, Percival M., *Diagnosing Personality and Conduct*, The Century Co., New York, 1931, p. 260.

³ Voelker, Paul F., *The Function of Ideals and Attitudes in Social Education*, Teachers College, Contributions to Education, No. 112, New York, 1921.

⁴ Fernald, G. G., "The Defective Delinquent Class Differentiating Tests," *American Journal of Insanity*, Vol. 68, 1912, pp. 524-594.

⁵ Hartshorne, H., and May, M. A., "Testing the Knowledge of Right and Wrong," *The Religious Education Association Monographs*, No. 1, 1927; Hartshorne, H., and May, M. A., *Studies in Deceit*, The Macmillan Co., New York, 1928; and Hartshorne, H., May, M. A., and Shuttlesworth, F. K., *Studies in the Organization of Character*, the Macmillan Co., New York, 1930.

1. *An Opposites Test*, a word knowledge test related to character and conduct matters.
2. *A Similarities or Cross-Out Test* (similar in purpose to No. 1).
3. *An Ethical-Social Vocabulary Test*.
4. *A Word Consequences Test*, in which the subject indicates, for an action given, *all likely* consequences, the *most likely* consequence, the *best* consequence, and the *worst* consequence.
5. *A Cause and Effect Test*, in true-false form, of which the following are samples:
 1. Good marks are chiefly a matter of luck *True False*
 2. Ministers' sons and deacons' daughters
usually go wrong *True False*
6. *A Duties Test*. The pupil marked an item *Yes* if he considered it his duty, *No* if he considered it not to be his duty, and *?* if sometimes his duty and sometimes not. Samples are:
 1. To help a slow or dull child with his lessons. *Yes ? No*
 2. To read the newspapers every day. *Yes ? No*
7. *Comprehension Test*, a multiple-choice type that tests "what one should do." Samples are:
 1. If someone asks to borrow your pencil,
----- Tell him it's broken
----- Tell him that you just lost it
----- Tell him that you don't want to lend it
----- Let him take it
 2. If someone steals your lunch
----- Steal another lunch to even it up
----- Report it to the teacher

- Cry about it
 ----- Say nothing about it

8. *Provocations Test.* This test represents an attempt to determine the extent to which judgment has the ascendancy over wishes, prejudices, and emotions. *R* encircled as an answer indicates the person taking the test thinks the action exactly right, *Wr* absolutely wrong, and *Ex* wrong but excusable under the circumstances. Samples are:

1. Helen noticed that nearly everyone in the class was cheating on a test, so she cheated too. *R Ex Wr*
2. Harry was a Christian boy. One day a Jewish boy called Harry "a dirty Christian." Harry knocked him down. *R Ex Wr*

9. *A Foresight Test.* Consequences of given situations are to be checked as "likely to happen," "might happen but not likely," or "would not happen." A sample is:

1. John accidentally broke a street lamp with a snowball. (Consequences to be checked in columns of squares labeled as above).
 - a. John was arrested and sentenced to six months in jail.
 - b. John said nothing about it, and people thought another boy had done it.
 - c. The emergency wagon had to come and repair it.
 - d. He thought it was such fun that he smashed a lot more lamps.
 - e. There was an accident there because it was dark.
 - f. Some people were cross about it, and John's father got into trouble.
 - g. The glass went on the street and a child cut his hands on it.
 - h. The city had to pay for the lamp.

10. *Recognitions Test.* A list of acts must be classified as Cheating (*C*); Lying (*L*); Stealing (*S*); Wrong, but not cheating, lying, or stealing (*X*); or Not wrong at all (*J*). Samples are:

1. Bullying younger children *C L S X J*
2. Using street-car transfers that are
out of date. *C L S X J*

11. *Applications Test.* This is a test of ability to apply principles. A sample will make its nature clearer:

1. Mary saw Helen cheating on an examination. She had to decide whether she would
☐ (a) Report it to the teacher.
☐ (b) Not report it to the teacher.

Here are the five rules, of which two apply to this problem. Check two and only two in the spaces at the left of the numbers.

- ☐ (1) Treat others as you would like to have them treat you.
- ☐ (2) Be true to what is for the good of all, even when your own interests or those of your friends are involved.
- ☐ (3) When you have wronged some one, ask to be forgiven.
- ☐ (4) Be cheerful and uncomplaining when disappointed or hurt or in trouble.
- ☐ (5) Do not think of yourself as more important than you are.

After checking the two rules that apply to Mary, put a check before either (a) or (b), according as you think it would have been right for her to tell or not to tell.

12. *A Test of Good Manners.* This test constructed by Miss C. I. Orr for the Character Education Inquiry measures knowledge of current standards of

courtesy and good manners. Items are of the *True-False*, multiple-choice, and *Yes-No* forms. Samples are:

1. In helping yourself to sugar always use
your own spoon. *True False*
2. When yawning, make no attempt to
suppress it by covering the mouth. . . . *True False*

How should we evaluate these tests? Are they really good tests of character, morality, and conduct? Should we recommend them for use in practical situations and put a great deal of dependence on the results of their administration? Hartshorne and May attempted to evaluate their tests on the basis of (a) reliability; (b) intercorrelations of the various tests, or relation with each other; (c) correlations of separate tests with a composite of several tests; (d) correlations of the tests with intelligence; (e) correlations of the tests with age; and (f) correlations of the tests with actual conduct (cheating). The tests in general possess satisfactory reliabilities, their reliabilities being only slightly less than one would expect for school achievement tests in similar form. So far as consistency of results is concerned, they can, therefore, be depended upon. The intercorrelations are mostly positive and of moderate degree, indicating some common factor running through the various tests. The correlations between separate parts and a composite of several of the tests are generally between .30 and .60. Performances on the tests among school children are generally positively related to intelligence and negatively related to age. Unfortunately, from the standpoint of definite recommendation of the tests for practical problems, they do not show high relationships with actual conduct, although the two are positively correlated. Perhaps we

should, however, demand further investigations along this line, since the studies have been limited and in many cases based upon rather inaccurate subjective estimates of conduct. Symonds' summary in discussing these tests seems a pertinent close to our consideration of them:

It is possible to measure knowledge and judgment with reference to conduct through the application of several useful tests which have been constructed for measuring health, Biblical knowledge, ethical knowledge, etc. These tests have very satisfactory reliability, comparing favorably with tests in the school subjects, similarly constructed. They correlate somewhat with each other and substantially with intelligence in general. The correlations with conduct are very low, so that with the less perfect tests they seem to fail to differentiate between normal and delinquent individuals. Conclusions from the research work done with these tests seem to indicate that answers reflect the code of the group in which the individual happens to be rather than any reasoned solution to the problem situation presented for judgment. These codes seem to be group affairs, and there is a distinct correlation between conduct and knowledge when groups as a whole are considered. The low correlations of knowledge and conduct for individuals indicate how distinct the two forms of activity are. On the other hand, when these correlations are compared with the correlations between different forms of conduct, there is ground for the suggestion that perhaps knowledge and judgment of conduct constitute after all the one force, however ineffective, that works toward integrating conduct.⁶

II. Behavior Tests

The Character Education Inquiry tests of honesty and trustworthiness will be described as examples of the be-

⁶Symonds, Percival M., *Diagnosing Personality and Conduct*, The Century Co., New York, 1931, p. 294.

havior or performance tests in character measurement. The battery included nine types of tests, as outlined below.

1. Testing dishonesty by the double testing technique.

The method consists in giving two forms of a test equated in difficulty, one form being given under conditions which do, and the other under conditions which do not, permit cheating. Variations in score exceeding the normal variation to be expected are considered evidence of dishonesty. The Character Education Inquiry made use of four tests in this group—an arithmetic test, a sentence-completion test, an information test, and a word-knowledge test. Cheating on one set of each was made possible by allowing the pupil to score his own paper by an answer key.

2. Speed tests. The technique is similar to that just described. In these tests, however, performance is determined by speed rather than power. The subject can cheat by adding extra items to the answers when allowed to score his own paper. Hartshorne and May utilized speed tests in addition, number-checking, letter cancellation, digit cancellation, digit-symbol substitution, and dot-in-square placement.

3. Coördination tests. These included tracing tests, such as tracing around squares, keeping within a narrow margin; and a circles test, as illustrated in Fig. 22.[†] Under the test conditions, performance must be done with the eyes closed. Cheating or dishonesty was evidenced by achievements beyond the limits of probability without "peeping." The limits of probability were determined by trials on control groups. The method of

[†] Reproduced by permission of the Macmillan Company, New York.

these tests is known as the *improbable achievement technique*.

4. **Puzzle tests.** The puzzle tests utilized in the Hartshorne-May studies were also based on the improbable achievement technique. The puzzles used, while appearing simple, could not be solved without cheating.

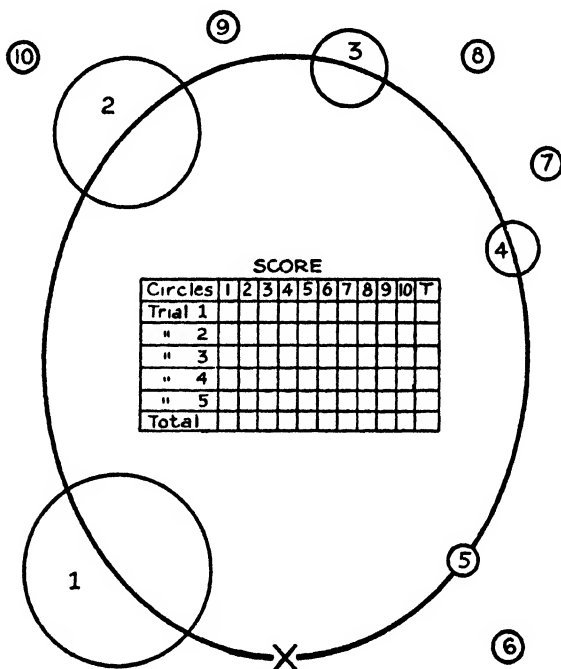


Fig. 22.—Circles Test for Dishonesty. (*Directions.*—Wait for the signal for each trial. Put the point of your pencil on the cross at the foot of the oval. Then, when the signal is given, shut your eyes and put the figure 1 in each circle, taking the circles in order. For the second trial, put the figure 2 in each circle. For the third trial, the figure 3, etc. After each trial put a check mark in the score box under the number of each circle you succeeded in striking, count the checks, and enter the total in the column headed T at the right of the score box. After the last trial add up Column T. This is your score. The maximum score is 50.)

Correct solutions, beyond certain limits allowable by chance, indicated dishonesty.

5. Lying tests. Two types of tests to discover lying were utilized in the studies we have been discussing. One test consisted of questions regarding tests taken previously (those described under No. 1). The questions asked directly whether the subject had cheated. Without the subject's knowledge, the truth of his answers could be checked. This test was considered to measure *lying to avoid disapproval*. Another type of test was used to measure *lying to gain approval*. This consisted of a number of questions to be answered *Yes* or *No*, which, although they have widespread social approval, are rarely actually done. Sample questions are:

Do you usually report the number of a car you see speeding?	Yes	No
Do you always obey your parents cheerfully and promptly?	Yes	No
Do you read the Bible every day?	Yes	No

A critical point, separating honesty from lying, was established by having adult graduate students answer the questions so that these truthfully represented their childhood. Children who received scores considerably above this critical point were judged to have lied to gain approval.

6. Homework tests of cheating. A form of the word-knowledge test described under (1) was given to be done at home. Instructions against obtaining help from the dictionary or another person were given twice. Scores considerably above the performance on the test in the classroom were considered evidence of cheating or dishonesty.

7. **Athletic contests.** Deceptive behavior was measured in these tests by the subject's inclination to fake a good record. The physical tests used were the dynamometer test, the spirometer test, the chinning or pull-up test, and the standing broad jump. Certain prizes were offered for good performances. Each child was allowed three trials, the best out of the three being noted by an examiner without the child's knowledge. The child was then encouraged to make five additional trials immediately after the three without any supervision, and to report them later to the examiner. Better records than in the first three are highly improbable, because practice effect is practically absent in the tasks used, and fatigue effect is considerable. Gains stated by the child above a normal variation indicated deception.

8. **Party games tests.** Cheating was observed in performances in such games as "pinning the tail on a donkey" and "bean relay race." The game situations were so arranged that cheating could occur and could be checked without the subject's knowledge.

9. **Money tests.** Situations were worked out in which was given the opportunity to take money, dishonestly, without apparent danger of detection. Systems of checking made the detection of dishonesty possible.

How do these various direct-method conduct measures stand the tests of a good measuring device? The studies by Hartshorne and May indicate that the tests are reliable as judged by correlations between original tests and retests. The tests, however, raise grave doubts regarding their general value on two counts. They have not been demonstrated to correlate to any close degree with general behavior; and they seem to test only very specific situations, or conduct only under very specific circumstances. The studies of the tests indicate that a single

test of deceit or a single test of honesty bears little relationship to deceit or honesty in general. The reader will probably be impressed by Symond's excellent summary of the value of such tests of conduct:

From a review of all the skillful and ingenious methods for testing conduct directly that have been devised, the conclusion stands out above all others that conduct is very specific. When exactly the same test is repeated, the correlation is fairly high, perhaps around .70 or .80. But when the situation is changed ever so slightly, the correlation between the two similar tests drops, and long before the two situations seem different enough to be called by different names, the correlation has dropped close to zero. A battery of tests designed to test such a trait as persistence, or aggressiveness, or speed of decision gives results so varying and with so little consistence as to furnish little warrant for assuming the presence of such a trait.

These low intercorrelations also help to explain the low correlations of these tests with outside criteria. Since the tests are so specific as to fail to correlate with tests bearing the same name, naturally they could not be expected to correlate to any degree with other factors which are admittedly dissimilar.

The conclusions that one draws from these results are not very encouraging. There are four possible things that may be done with performance tests in the measurement of conduct:

1. They may be discarded as being so specific as to be useless for all practical purposes.

2. Tests may be devised that apply to the specific situation in which they will be used. Since tests are so very specific, test situations must be set up which approximate as closely as possible the situations in office, industry, school, or institution where they will be used.

3. Since no one test measures a given quality adequately, a variety of tests representing a range of situations in which the trait occurs may be devised so that the

composite will be a satisfactory measure of the trait in question. . . .

4. A fourth method of using these tests is to pick out one test for each of a number of different traits and weigh these tests in combination in order best to predict success in business or school. . . .

In all of these alternative methods of using performance tests certain practical problems of cost, difficulty of administration, difficulty in applying statistical techniques, and the like, arise which very definitely limit the use of these tests. It is often expensive to test in the practical situation. Certain ingenious devices must be applied which eat into time and money, as May and Hartshorne found out. Again, to give a well-rounded battery of tests is also expensive. Finally, one who plans to use the regression equation technique must count the cost beforehand.

Performance tests have a real and valuable place at the present time in experimental work. As used by Hartshorne and May, they have revealed facts that were obtainable in no other way. But there must be considerable further development before tests of this type become a feasible tool in clinical work.⁸

III. Other Types of Tests

Although the tests we have described represent the best and most useful of the tests that have been devised for measuring character, they do not constitute a complete list of types of character measures. The Commission on Character Education of the National Education Association lists eight types of tests and measures that have been used in character measurement.⁹ These are (1) *Measures based on physical factors*; (2) *Measures based on significant knowledge*; (3) *Measures of opinion*;

⁸ Symonds, Percival M., *op. cit.*, pp. 352-354.

⁹ *Character Education*, Tenth Yearbook, Department of Superintendence, National Education Association, Washington, D. C., 1932.

(4) *Self-descriptive measures*; (5) *Disguised measures*; (6) *Measures of conduct in controlled situations*; (7) *Significant facts as character measures*; and (8) *Reputation measures*. The second and sixth of these designate the types of tests we have described from the Character Education Inquiry tests. The fourth and fifth represent adaptations of personality tests, such as are described in Chapter XXI, to the study of character traits. The last two, if more than rough, uncontrolled personal estimates, represent the application of rating-scale technique to evaluating character and reputation.

Measures based on physical factors, as discussed by the Commission on Character Education, include (a) body type, (b) appearance, (c) motor reactions, as handwriting, (d) biochemical tests, and (e) physiological changes during emotion. The first three of these are generally ancient attempts to tell character which have not stood the tests of validity, and are now relegated to the pseudo-scientific. Biochemical methods offer hope, and are certainly worthy of investigation. For the most part, at present, the chemical processes of the body are not well enough understood and easily enough measured to be useful psychologically. We know that endocrine secretions do influence basic attitudes and conduct, although no specific form of immoral or antisocial conduct is traceable to any particular gland. We know, also, that certain infections, toxins, and drugs are clear-cut causes of personality disorders. Fairly recent studies suggest relationships between personality and character differences and such biochemical factors as salivary alkalinity, urinary acidity, creatinine in urine, blood phosphorus and blood calcium. Further research may eventually prove these biochemical tests to be of practical value in judging phases of character.

Physiological changes during emotion have been utilized to indicate the emotional effect of confusion and hence to detect attempts to deceive or lie. The physiological changes studied include blood pressure, breathing, reaction time, and the psychogalvanic reflex. These are discussed and evaluated in Chapter XXVI.

Measures of opinion and attitude, particularly toward matters of social welfare, law abidance, and morality, should certainly have a bearing upon character testing. One's opinion or attitude is very likely to color one's conduct; in many instances it is likely to be the most important motivating force in conduct. Thurstone of the University of Chicago has embarked on a program of working out and statistically evaluating a considerable number of attitude scales. Scales published or in preparation include, among a number of others, scales for measuring attitude toward God, the church, birth control, movies, law, Sunday observance, criminals, capital punishment, and freedom of speech. Each scale consists of a list of statements of the "pro" and "con" point of view, with reference to the subject of the scale. The person filling out the scale checks those statements with which he agrees. Each statement has a scale value indicating the extent to which the statement is favorable toward the subject of the scale. Attitude or opinion is indicated by the sum of scale values of those statements checked. For the method of evaluating the statements of the scales, the reader is referred to the original writings. The scope of this book does not allow an elaboration on this.¹⁰

The validity of attitude and opinion scales is directly related to factors which insure an honest expression in

¹⁰ Thurstone, L. L., and Chave, E. J., *The Measurement of Attitude*, University of Chicago Press, Chicago, Ill., 1929.

the subject's answers. Opinion tests are of most value when there is an important and genuine motive leading the persons being tested to give a true picture of their opinions. The procedure of giving the tests, including the manner of introducing them and explaining them, is as important in affecting the worth of the test as what is included in the test items themselves. A discussion by Thurstone is interesting in this connection:

There comes to mind the uncertainty of using an opinion as an index of attitude. The man may be a liar. If he is not intentionally misrepresenting his real attitude on a disputed question, he may nevertheless modify the expression of it for reasons of courtesy, especially in those situations in which frank expression of attitude may not be well received. This has led to the suggestion that a man's action is a safer index of his attitude than what he says. But his actions may also be distortions of his attitude. A politician extends friendship and hospitality in overt action while hiding an attitude that he expresses more truthfully to an intimate friend. Neither his opinions nor his overt acts constitute in any sense an infallible guide to the subjective inclinations and preferences that constitute his attitude. Therefore, we must remain content to use opinions, or other forms of action, merely as indices of attitude. It must be recognized that there is a discrepancy, some error of measurement as it were, between the opinion or overt action that we use as an index and the attitude that we infer from such an index.¹¹

¹¹ Thurstone, L. L., "Attitudes Can Be Measured," *American Journal of Sociology*, Vol. XXXIII, No. 4.

PART VII

PHYSIOLOGICAL MEASUREMENTS
IN PSYCHOLOGY

CHAPTER XXIII

Measurement of Fatigue

ALMOST every example of sustained production or performance involves the problem of fatigue. The applied psychologist has met the problem *in the workshop*, where fatigue has been the factor accounting for decreased production, occurrence of accidents, or adverse working attitude; *in the office*, where fatigue has limited accuracy of performance or has called for unnecessary energy expenditure to maintain work at the required efficiency; and *in the classroom*, where fatigue has influenced learning ability. More recently, fatigue studies of automobile drivers have been directed toward solving the problem of *street and highway accident prevention*. Finally, fatigue studies have interested *the clinical psychologist* because they throw light upon the basic causes of neurasthenic or similar mental conditions.

All of these studies have emphasized the need for quantitative measures of fatigue. We can expect relatively little to be accomplished in increasing production by eliminating causes of fatigue until quantitative studies of fatigue can be made. Little can be learned about the relationship between accidents and fatigue until there are available quantitative indications of degree of fatigue. It is difficult to decide upon the importance of fatigue as a factor in mental disorders and disease without means of indicating the amount of fatigue that has accumulated.

The means of measuring fatigue which we find in psychological literature can generally be grouped under three methods. The earliest psychological methods measured fatigue *in terms of decreased production* or decreased performance. The various studies depicting "fatigue curves" throughout a day of work or a day of activity are of this type. These measurements err in assuming that production is inversely proportional to fatigue. In many instances this is not true. Other mental or emotional factors often interfere with production when actual fatigue as measured by more accurate methods is slight. Such measurements are likely to contribute very little to the solving of the problems in which fatigue is a factor.

A second method of measuring fatigue, which is met in a few of the psychological studies, is subjective; fatigue is measured by a *subjective feeling of tiredness*, recorded by the subjects being studied. Such a method of measuring fatigue, in the first place, is open to all the objections pertaining to subjective measures in general, and in addition it is open to the objections just mentioned in connection with measurements in terms of production. Subjective feelings of tiredness are not by any means perfectly correlated with actual fatigue. They are not even always related to amount of production. Poffenberger demonstrated the lack of correlation between feelings of tiredness and output in an experiment in which twelve subjects did mental work continuously for five and one-half hours, indicating at intervals the quality of their feelings on a scale ranging from "extremely good" to "extremely tired." At these same intervals records were taken of their production. The relationship between the two is shown in Fig. 23.

The third type of measurement of fatigue is one in

terms of *physiological state* of the individual. This method seems to possess many advantages over the other two, in that it can be made objective, it is directly related to the nature of the fatigue itself, and it is independent of the production or accomplishment which usually constitutes the problem for which fatigue is studied.

Further back than we can find written records, certain physiological effects of fatigue have been noted. Such

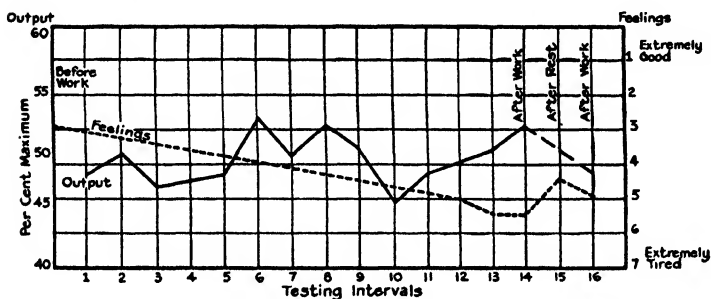


Fig. 23.—The Relationship Between Production and Feelings of Tiredness.

obvious manifestations of strenuous physical exercise as increased breathing, increased heartbeat, and increased perspiration could not fail to be noticed. The exercise, however, has to be rather strenuous for such changes to be observable without more exact instrumental measurement. Present problems demand quantitative measurements and often the detection of changes produced under circumstances of rather mild exercise.

Physiological studies of the blood under various conditions led, long ago, to a theory of fatigue as a condition produced essentially by an accumulation of fatigue products in the blood, such an explanation being credited by experiments demonstrating that a rested animal can be

fatigued by injecting into its blood stream the blood from a fatigued animal. Fatigue is to be looked upon as a general body condition in which observation of the following may be stressed: the *muscles*, as the primary seat of energy transformations; the *blood*, as a vehicle of energy supplies and waste removal; the *heart and blood vessels*, as the distributing mechanism; the *respiratory, alimentary, and excretory systems*, as sources of supply and means of waste removal; and the endocrine glands and *nervous system*, as the coördinating and regulating mechanisms. Quantitative measurements of fatigue of the physiological type are based upon changes in these parts of the body. Not all the changes can be measured with our present devices, and not all the measurements that can be made are practical for ordinary psychological application, because of their complexity and technical nature or, in some instances, because of the discomfort which their application may cause the subject being measured.

In the remainder of this chapter are indicated those means of measuring fatigue which have proved of most value and are most practicable. The discussion is based upon the results of an experimental investigation conducted by the Psychology Department at the George Washington University.¹ The study aimed primarily to demonstrate the feasibility of utilizing certain physiological measurements in denoting the amount of fatigue produced in an individual. The subjects used in the study were university men between the ages of 18 and 35. "Normal" physiological measurements were made on each subject before the experiment. The subjects

¹ Moss, F. A., Roe, J. H., Hunter, O. B., French, L., and Hunt, T., "Measurement of Fatigue by Physiological Methods," *Journal of Experimental Psychology*, Vol. XIV, No. 4, Aug. 1931.

then engaged in a fatigue-producing occupation, and the same physiological measurements were made after fatigue had set in, in some instances the measurements being taken at varying stages of the fatigue.

The method of producing fatigue consisted in riding a bicycle ergometer. This is a machine originally used by Benedict and described by him as follows:

The rear wheel of a bicycle was replaced by a copper disk 40.5 centimeters in diameter and 6 millimeters thick. This disk is mounted in such a way that it rotates freely on a ball-bearing axle. A small sprocket wheel is attached to the axle and is in turn connected in the usual manner with the large pedal sprocket wheel by means of a sprocket chain. A wooden frame surrounds the periphery of the disk, and to the upper part of the frame is attached an electro-magnet. Binding posts are attached to the magnet to connect with the electric cable leading to the observer's table, where strength of current can be regulated with great accuracy. The field of the magnet is so extended that the copper disk rotates in the center of the field with but a very small air gap between the surface of the disk and the surface of the magnet, and hence the resistance is wholly that of magnetic induction. A current of 1.25 amperes induces large eddy currents in the copper disk to such an extent that the resistance is very noticeable.²

By varying the strength of the magnetic field, regulated through a rheostat and ammeter, the amount of work performed in pedalling can be varied. The number of revolutions made is measured by an electric counter connected with the sprocket wheel. The ergometer has been calibrated in terms of calories, the work performed per revolution being .0231 with a current of 1.25 am-

² Benedict, Francis G., and Carpenter, Thorne M., *The Influence of Muscular and Mental Work on Metabolism*, United States Department of Agriculture, Bulletin No. 208, 1909, Washington, D. C.

peres and 70 revolutions per minute, this being the speed and amperage used in the present experiment. During the experiment, each subject rode the bicycle for 15 or 20 minutes, steadily, at the rate indicated.

1. **Blood pressure.** Blood pressure is related to the force with which the circulatory system is supplying blood to the various parts of the body. Since activity, exercise, and fatigue put extra demands upon the circu-

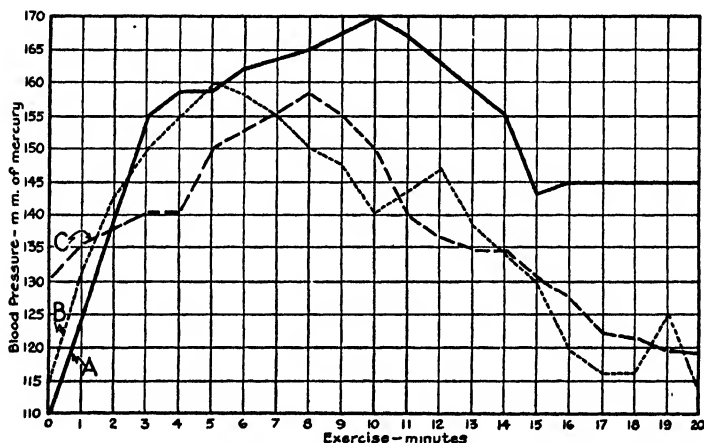


Fig. 24.—Blood Pressure Curves in Fatigue (Three Subjects).

latory system, we might expect blood pressure to be proportional to the demands, or in other words, to degree of fatigue. Blood pressure measurements in the experiment were made using an ordinary sphygmomanometer. During 20 minutes' exercise on the bicycle, blood pressure curves were obtained like those shown in Fig. 24. Up to about 8 or 10 minutes of exercise, the subjects showed an increasing blood pressure, with increasing exercise. Beyond this point, blood pressure decreased until at the end of 20 minutes, for many of the subjects, it had approached the initial pre-exercise level. The

blood pressure curves generally show a rapid rise to a maximum height, then a gradual fall. Onset of fatigue seemed to follow the downward slope of the curve. The more enduring subjects showed maximum height later.

2. Carbon-dioxide combining power of the blood. Under normal conditions the human blood contains a reserve supply of fixed alkalis or bases. It is this "alkali reserve" which keeps the hydrogen ion concentration, or true reaction of the blood, slightly alkaline. In processes of metabolism in the body an excess of acid radicles, from such things as the combustion of carbon substances, is continuously produced. During exercise there is a marked increase in these acid radicles, though in normal subjects they are not sufficient to make the reaction of the blood acid. This condition, however, can be measured by the carbon-dioxide combining power of the blood, which is lowered by depletion of the supply of fixed bases. Since the fixed bases of the blood constitute the chief means of transporting carbon dioxide from the tissues to the lungs, depletion of the supply of fixed bases reduces the capacity of the blood to carry carbon dioxide. This condition so produced leads to an accumulation of carbon dioxide in the tissues and consequent blocking of the processes of oxidation, factors important in producing the condition of fatigue.

3. Blood sugar. Blood sugar constitutes the fuel material for activity and energy production. During periods of demand it is liberated in relatively large quantities from its storage place, principally the liver. As activity continues, its supply in the body becomes less and less. Blood sugar records for the subjects studied showed considerable variation, dependent upon the stage of fatigue at which measurement was made. Increased amounts of blood sugar were found in the initial stages of fatigue or exercise, but decreased amounts were found

at later periods, when it can be assumed that the extra liberation of blood sugar in the body had been more than offset by the extra demand for it. Equal exercise does not affect everyone equally with respect to blood sugar changes; and since individual differences in blood sugar reactions are wide, such records seem hardly practical as an index or measurement of fatigue.

4. Metabolism. The term "metabolism" usually refers to the sum total of energy changes going on in the body. Since energy changes going on in the body are increased in proportion to the amount of exercise or work being done, it might be supposed that measures of metabolic rate as compared with resting rate would constitute good indications of amount of fatigue. Metabolism changes in the George Washington experiment were measured by a Sanborn graphic metabolism machine. Measurements of "basal metabolic rate" showed an average increase of 54.5 per cent. Fig. 25 shows a typical tracing before and after fatigue. Note in the lower record, after fatigue, the longer strokes, signifying deeper breathing; the greater number of strokes, signifying more frequent inhalations; and the greater slope of the curve, indicating increased consumption of oxygen.

5. Blood cell studies. The red blood cells, particularly their hemoglobin content, are concerned in the carrying of oxygen, which must be supplied in extra quantities during and immediately following processes which produce fatigue. White blood cells exhibit a protective reaction in increased numbers whenever there are injurious or toxic substances to be eliminated from the body. Since injurious waste products seem to be the most important causal factor of fatigue, it would seem logical to find these cells increased in proportion to the fatigue. The experimental study showed an average

increase in red blood cells of about 6 per cent and of white blood cells of about 57 per cent.

The various physiological changes produced by fatigue are summarized in Fig. 26. The two types of measurement that seem to offer the most reliable quantitative indications of fatigue are measurement of metabolism

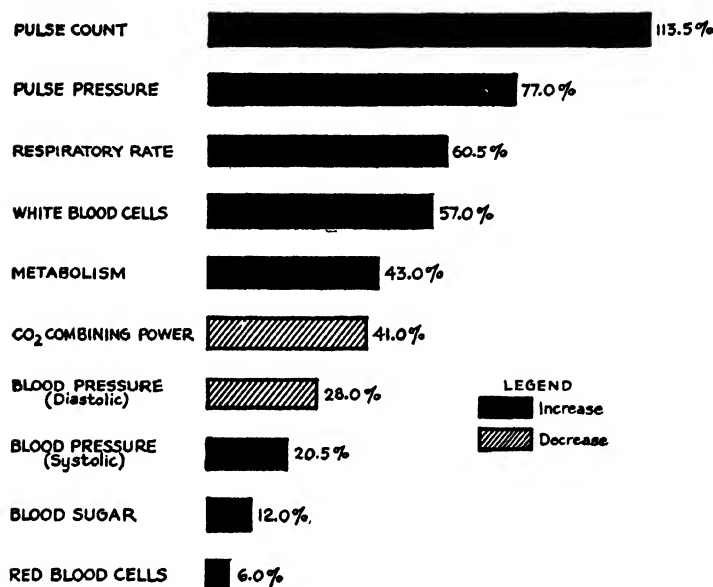


Fig. 26.—Physiological Changes Produced by Fatigue.

and measurement of carbon-dioxide combining power of the blood. Both of these seem to be sensitive to relatively slight amounts of fatigue, and both are relatively independent of transitory mental and emotional causations for their change, which make some of the other methods, as that for blood pressure, rather impractical. In the utilization of any of these physiological measurements of fatigue, the desirability of obtaining normal measurements for subjects to be studied is clear.

CHAPTER XXIV

Laboratory Tests in Mental Disorders

WE MIGHT expect to find the basic differentiation between the mentally disordered or insane person and the normal person in the chemistry of their bodies. For certain types of disorders this is possible. The hallucinated, tremulous alcoholic has a system saturated with a toxic chemical; the nervous, irritable hyperthyroid is the result of too-powerful action of the chemical from the thyroid gland; the stupid, poorly developed cretin is the result of too little of this same glandular chemical; the silly, grandiose parietic patient has his insane characteristics because of an infectious agent that has invaded his blood stream and nervous system; the neurasthenic or neurotic individual might be defined in terms of the accumulated chemical waste products of an overactive life. All these examples of mental disorders, and many others, furnish psychological problems which can be really understood only by the utilization of physiological measurements. Whatever information may be gained through quantitative studies of intelligence and personality, in which the measuring instruments we have already discussed are utilized, should be supplemented by whatever physiological tests are available. It is only to be regretted that these tests cannot be extended to many of the so-called "functional" mental disorders, as yet so little known from the standpoint of their chemistry.

In the subsequent paragraphs we shall describe some

404. Laboratory Tests in Mental Disorders

of the more important tests that have been found of service in the field of mental disorders or abnormal psychology.

I. Blood Tests

Should we try to observe one thing that would give us the most information about the body as a whole, we should select the blood stream. The blood stream carries chemicals which nourish all the cells of the body; it carries the oxygen necessary for combustion of the body fuel; it carries the enzymes that stimulate the digestive processes in the body; it carries the endocrine products which speed up or slow down the body processes; it carries various waste products which accumulate in the body from fatigue; and, lastly, it bears the burden of many toxic products which either gain entrance into the body from the outside or arise from the body itself. It is not surprising, then, that we should examine the blood in an attempt to disclose the basis of a mental disorder. The more important blood tests made in connection with mental disorders are described below.

1. **Tests for syphilitic infection.** The two tests commonly used for discovering syphilitic infection are the Wassermann reaction and the Kahn tests. When an individual has syphilis, the entrance of the invading syphilitic germs causes the body to build up in the blood stream certain counteracting substances of a protective nature in an attempt to rid the body of the invading organisms. The laboratory tests for syphilis are, in principle, tests for discovering the presence of these counteracting bodies in the blood stream.

The importance of these tests lies in their diagnostic value in paresis and *tabes dorsalis*. The former is a mental disorder likely to be accompanied by marked mental

deterioration and grandiose delusions, and the latter is a disorder showing many nervous symptoms. Some 95 per cent or more of paretics show a positive reaction in the blood. Between 50 and 75 per cent of *tabes dorsalis* cases yield a positive blood Wassermann. It should be remembered that a case with a negative blood Wassermann may show a positive reaction in the spinal fluid. Hence the importance of the spinal fluid test in nervous cases.

2. Blood cell tests. These include tests to determine the number of red blood cells per unit volume of blood; the amount of hemoglobin in the red cells; and the number of white cells per unit volume of blood. The first two of these are important primarily because of the importance of the red blood cells in carrying oxygen to the various parts of the body. We might expect to find mental as well as "physical" disturbances when the efficiency of this transportation of oxygen is interfered with by a deficiency of red cells or by low hemoglobin content. White cell measurements are of importance primarily in infectious disorders, and since a considerable number of mental disturbances are infectious in origin, white cells are often of importance in diagnosing and dealing with mental abnormalities.

We may quote one example of such mental disturbances:

A boy, age 13, who had previously been getting along very well in school suddenly began to have trouble with his work. He found it difficult to go to sleep because he constantly heard voices which taunted him and said mean things to him. He had been hearing these voices for several weeks when first examined. On examination he was found to be introverted and emotionally flattened, apparently living in a world of his own creation. The blood tests gave him a hemoglobin of 48, and a red

406 Laboratory Tests in Mental Disorders

cell count of 2,100,000 (5,000,000 are normal). He was put on heavy doses of liver extract to build up his blood condition, and as the hemoglobin gradually increased, his mental condition improved. After the hemoglobin reached 70 no "voices" were heard. His school work improved, and he passed without difficulty to the next grade. It has now been almost a year and he shows no signs of a relapse.¹

3. Blood sugar. Chemical analysis of the blood to determine the amount of the blood sugar is of particular importance in diabetes. Since mental symptoms and coma often result from diabetes, the blood sugar test can be considered as having some value in the interpretation of mental disturbances. Recent studies of blood sugar curves, after the subject has been fed quantities of glucose, indicate that perhaps the bodily reaction to increased sugar varies according to types of mental symptoms.²

4. Carbon-dioxide combining power. Laboratory tests are available for determining the capacity of the blood for carrying carbon dioxide. Such tests are indicative of the state of fatigue or exhaustion existing in the body. They may be of value in diagnosing and suggesting treatment for certain cases of neurasthenia.

II. Spinal Fluid Tests

The whole nervous system is bathed in a fluid which serves primarily to protect it from shocks. Since this fluid is in such intimate relation with the nerve tissue, it is not surprising that we should find reflected in its

¹ Moss, F. A., and Hunt, T., *Foundations of Abnormal Psychology*, p. 390, Prentice-Hall, Inc., New York, 1932.

² Drury and Farran-Ridge, "Types of Blood Sugar Curves Found in Different Forms of Insanity," *Journal of Mental Science*, Vol. 71 (1925), p. 12.

chemical composition disturbances that take place in the nervous system. For spinal fluid tests a small amount of spinal fluid is withdrawn through a puncture made in the lumbar region of the spinal column.

1. **Wassermann reaction.** This is a test of syphilitic infection in the spinal fluid similar to the blood syphilitic tests already described.

2. **Lange's colloidal gold test.** In a colloidal gold test, ten test tubes are arranged in a row. Each test tube contains five cubic centimeters of colloidal gold solution

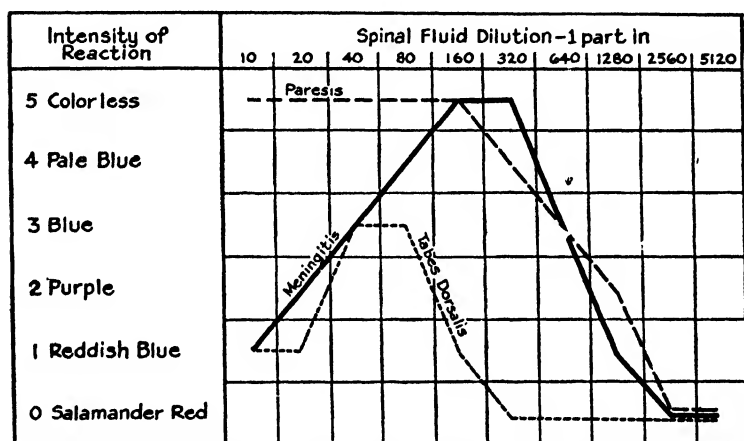


Fig. 27.—Colloidal Gold Curves in Various Mental Disorders.

and one cubic centimeter of spinal fluid diluted with physiological salt solution, the dilution ranging from 1:10 in the first tube to 1:5120 in the tenth tube. After the tubes have stood for twenty-four hours, readings for all ten tubes are taken. The intensity of reactions in the successive tubes is determined by color. Normal spinal fluid shows the original salamander red color in all tubes. The negative reaction, or zero reading, is salamander red in color, as in the beginning; a slightly positive reaction, or reading of 1, is reddish blue; the next,

408 Laboratory Tests in Mental Disorders

a reading of 3, is blue; the next, a reading of 4, is pale blue; and the strongest intensity of reaction, marked by complete precipitation of the colloidal gold, reading 5, is colorless. The readings from the test tubes are commonly shown by colloidal gold curves. Fig. 27 shows the difference in types of curves obtained in three typical mental disorders.

3. **Cell count for spinal fluid.** Increase in the number of cells in the spinal fluid is found in practically all cases of paresis and cerebral syphilis, in the majority of cases of lethargic encephalitis in the acute stages, and in acute infections of the meninges. Red cells, indicating blood in the spinal fluid, may be found in cases of traumatic injury to the brain or spinal cord, tumors, and hemorrhages. Normal spinal fluid contains a few white cells but no red cells.

4. **Sugar.** Chemical analysis of spinal fluid for sugar is sometimes made, since in cases of sleeping sickness and dementia praecox there is often an increased sugar content in the cerebrospinal fluid.

III. Basal Metabolism

Metabolism tests measure the speed of chemical reactions which go on in the living cells in connection with the immediate maintenance of the living state. It is usually measured by the amount of oxygen which the person consumes in a resting state after a fast of at least six hours. Expressed as a *basal metabolic rate*, it is the ratio of the amount of oxygen actually consumed in a unit time compared with the amount which the person should normally use according to his age, height, and weight. High metabolic rates are expressed as plus values over 0, as +18; low metabolic rates as minus values below 0, as -21.

In mental disorders, measurements of basal metabolic rates are particularly important in the diagnosis of thyroid disturbances. Low metabolism is considered practically diagnostic of thyroid insufficiency; and high metabolism of hyper-functioning of the thyroid.

IV. Blood Pressure

Blood pressure represents a reaction between the walls of the blood vessels and blood in the vessels. The maintenance of normal blood pressure depends primarily upon elasticity of the walls of the blood vessels and the integrity of the heart muscle. Abnormally high blood pressure usually means hardening of the blood vessel walls. In old people there is a marked tendency for loss of elasticity of these walls, with consequent rise in blood pressure. In conditions of arteriosclerosis, the hardening becomes quite marked and the pressure extremely high. Arteriosclerotic conditions in the blood vessels of the brain are invariably accompanied by mental symptoms.

Low blood pressure associated with an inefficient circulatory system may be of diagnostic value in some mental disorders. Such a condition is very commonly found in neurasthenia and frequently occurs in dementia praecox.

CHAPTER XXV

Glandular Function Tests

I. Nature of the Endocrine Glands and Their Importance to Psychology

THE endocrine glands are relative newcomers in the field of psychological measurement. While their existence has been known for a considerable time, quantitative measurements in the field of gland study have had to await years of physiological and chemical research into the nature and effect of the glandular secretions. Even now this research should probably be regarded as just at its beginning. We shall find that in most instances the quantitative measurements have reached only a very crude state of development.

The endocrine glands are relatively small collections of specialized cells located at scattered places in the body. They exert their influences by manufacturing and pouring into the blood stream chemical substances (endocrines, internal secretions, or hormones) which, through their effect on various parts of the body, markedly influence the development and behavior of the individual. The hormones have often been called chemical regulators of the body. In this capacity they preside over four main functions: (1) growth and development, including both physiological and mental growth; (2) sex and reproduction; (3) nutrition and general metabolism; and (4) secretion from other glands, including maintenance of proper glandular balance with reference to the secretions of all the glands.

Chief among the endocrine glands of interest to the psychologist are the thyroid, the adrenals, the gonads (sex glands), the pituitary, the pancreas, and the parathyroids.

The *thyroid gland*, which encircles the windpipe, is the chief regulator of metabolism. It is the chief determinant of the rate at which the various physiological and psychological functions of the body take place. It has often been termed the gland which regulates our "speed of living."

The *adrenal glands*, which lie just above the kidneys, produce two major hormones, adrenalin and cortin. Adrenalin is the hormone liberated in increased amounts in emotional states, and in other states of extreme exertion. Cortin is a hormone necessary to the maintenance of circulatory efficiency, and in turn to the maintenance of life. Deficiencies lead to circulatory collapse, weakness, and finally death. Overproduction of cortin has been noted by some to influence sex characteristics. Tendencies toward masculinity in the female are sometimes based upon excessive adrenal function.

The *gonads* profoundly influence man both psychologically and biologically. They determine in part the development of the various secondary sex characteristics, such as voice, bodily contours, size, emotional differences; in short, those things which determine the characteristics of masculinity and femininity. Biologically, endocrine secretions from the sex glands have many important roles to play in the process of reproduction.

The *pituitary gland*, situated at the base of the brain, produces several different hormones. One of these hormones controls skeletal growth. Excessive functioning may produce gigantism; deficient functioning, dwarfism. Circus giants and midgets are most frequently the result of such dysfunctions. Other important hormones from

the pituitary gland play a role in the sex cycle, and are essential to the normal functioning of sex glands and to the normal development of those traits characterizing the two sexes.

The *pancreas* is a gland only part of which is endocrine in nature. Its endocrine function is in relation to carbohydrate utilization in the body. Deficiency in this function produces a disturbance (*diabetes mellitus*) in which there is a marked rise in the sugar content in the blood. This condition is not infrequently accompanied by profound symptoms in the mental sphere.

The *parathyroid glands* have to do primarily with the metabolism and utilization of calcium in the body. Their mental and psychological relationships depend on the fact that disturbances in the calcium level are likely to be reflected in disturbances in nervous system sensitivity and irritability. Low blood calcium is likely to produce increased neural sensitivity, which is manifested in conditions of tetany and convulsion.

The scope of this book does not permit a detailed discussion of the function of the various endocrine glands. This brief résumé of a few of them has been given merely to suggest their importance to those interested in human behavior.

The psychologist is interested in endocrine glands, and in measurements of their level of activity and efficiency of function, primarily because of their relationship to (1) mental development, (2) personality, (3) emotional reactions, and (4) mental disorders. An example of endocrine influence on *mental development* is furnished by cretinism. This is a condition caused by complete absence or defective functioning of the thyroid gland, occurring either congenitally or in early childhood. It occurs sporadically all over the world and is of very fre-

quent occurrence in certain districts where there is a deficiency of iodine—the chemical necessary for manufacture of thyroid secretion in the body. The symptoms of cretinism are usually noticeable during the first year of the child's life. Since the thyroid gland has the main function of presiding over general bodily metabolism, of controlling the speed of the chemical processes taking place in the body, these processes in the cretin go on at such a low ebb that he simply remains at an infantile level of development. Physically, his face retains the flattened features and broad expanse between the eyes characteristic at birth. He does not grow in stature and his growth is likely to be ill-proportioned. The legs and arms remain short, the abdomen protrudes and assumes a size out of proportion to the rest of the body. Mentally, the cretin is very much retarded. He makes few intelligent responses to his environment. He is unable to learn those things which the normal child learns. He does not begin to talk at the normal age, and his general adjustment to his surroundings is markedly below par. If untreated, he is likely to live his life at the level of an idiot or an imbecile.

Almost all glandular disturbances produce some effect upon *personality*. Among the cases short of actual mental disorder, we may cite the hyperthyroid as an example. These individuals are likely to have a personality colored by restlessness, by a sort of anxiety, and by a hyperexcitability to emotional stimuli. Their emotional control is usually somewhat defective and they are easily aroused to fits of temper. They often possess an extreme degree of "pep and energy," which may manifest itself in nervousness unless constantly directed toward the accomplishment of some task.

Another outstanding example of endocrine influence

upon personality is furnished by the condition usually referred to as Fröhlich's syndrome, which results from inadequacy of function of the pituitary gland in its relation to sex gland function. Clendening has given a graphic picture of this condition:

It begins before adolescence and occurs mostly in males, though it is found in females. The boys are fat, feminine, weak, and misunderstood. Their manifest deformities are regarded by their parents, teachers, and playmates as natural and inevitable variations of human structure, rather different, but within the normal limits. They are sissy because they are sissy: some boys are sissy. They are fat because some people are fat. They are weak and do not play boys' games because they are sissy. That is the usual view. It is not commonly recognized that they are definitely in a mutual deficiency group, that the deficiency is an affair of internal secretion, and that treatment to be effective at all must begin early in life.¹

As an example of endocrine function in *emotional states* we have already discussed in Chapter XXI the role of the adrenal glands. These have an important function to play in the production of the physiological state which is characteristic of all strong emotions. The thyroid gland also is related to emotional functioning. Long periods of emotional excitement tend to produce hyperthyroidism with various attendant symptoms. During the World War, for example, there was a striking increase in the number of cases of this type, presumably to be accounted for by the great emotional stress to which groups of people were subjected.

Among the *mental disorders* or insanities of endocrine origin we may note one of profound depression—*involu-*

¹ Clendening, L., *Modern Methods of Treatment*, C. V. Mosby Co., St. Louis, 1928, p. 232.

tion melancholia. This is a disorder occurring at the period of involution, when there is a regression in sex gland function with a subsequent general glandular imbalance. A few who are unable to weather the storm of this glandular readjustment and its effects on the nervous system become afflicted by an extreme depression. They develop a profound melancholia which someone has characterized as a "saturated solution of grief." They are continually beset with anxiety, worry, and despair, which have little relation to the actual circumstances of their existence. They often pace the floor for hours wringing their hands and moaning continually. Behavior of a voluntary, directed sort is at a standstill. They are unable to bring themselves to accomplish even simple tasks that need to be done. They are often negativistic toward doing the biddings of others, this negativism often being carried so far as to refuse food over long periods of time. Delusions, or false beliefs, are frequent—the commonest types are beliefs of sinfulness or unworthiness. Hypochondriacal delusions or absurd beliefs about one's body are also of frequent occurrence.

These examples of glandular influences on our psychological makeup are only a few of many that might be cited. *Mentality and mental growth*, while affected most markedly by the thyroid gland, may be affected less or more indirectly by other glands. *Personality* is affected probably by every endocrine gland in the body and, furthermore, by the particular balance in which the various glands are maintained in their relationship to each other. Some students of personality have even gone so far as to attempt to label personality types according to particular gland-function dominances. While much of this classification is unsupported by evidence, we need not discard the general idea back of it. *Emotions* may

be most often discussed in relation to the adrenal glands, but the difference between the emotionality of the hyperthyroid and that of the hypothyroid must suggest also the importance of this gland, and in the emotions of love and sex life the sex and the other glands involved in the sex cycle are of prime importance. Finally, as etiological agents in *mental disease*, the endocrines are involved in many of the psychoses. There are the thyroid psychoses associated with toxic conditions of the thyroid; epilepsies of endocrine origin; manic depressives and dementia praecoxes of glandular nature; and neurasthenias of pituitary dysfunction.

These examples in themselves will suggest the immense importance to the psychologist of studies of endocrine function. They are related not only to the specific aspects of behavior constituting the four categories which we have mentioned, but also to the problems of school progress, social adjustment, vocational guidance, and industrial and vocational efficiency.

II. Measurement in Relation to Endocrine Function

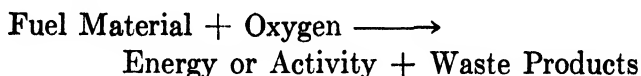
So far as quantitative measurement is concerned, we can readily appreciate the essential nature of a method to determine the precise degree of glandular activity of each of the endocrine organs. Unfortunately such methods are not generally available at the present time. When the different hormones have been chemically identified and the normal quantity of each in the blood has been analyzed, we can expect to have at our disposal a quantitative index by which to gauge the level of their activity. So far, only very few of the endocrine secretions have been chemically analyzed, and even for these there is no direct quantitative test which has proved practical. At present the level of functioning of the

endocrine glands is generally studied by indirect methods, and our quantitative measurements of their functions are usually in terms of their effect upon some part of the human body or upon some aspect of animal behavior. Some of the tests have reached the specificity of measuring the effect of the glandular secretion upon a specific part of the bodily function which seems to be controlled by the gland. This is illustrated by the measurement of insulin production by the pancreas in terms of blood-sugar level. The insulin of the pancreas is the important controlling factor in blood-sugar level; hence, we measure amount of insulin in terms of the sugar content in the blood. Others of the quantitative measurements of glandular function are in terms of much larger effects in the body or in terms of more general results of the endocrine function. An example in this category is the measurement of thyroid function in terms of its effect upon general body metabolism. These general measurements are more likely to be affected by other factors, and the measurements can usually be taken as indicative of glandular function only if the conditions under which the measurement was made are rigidly controlled. Finally, there is a considerable group of measurements of glandular function based upon the effect upon test animals of fluids of the body which contain the endocrine products. Most of the tests of sex hormones and pituitary hormones are based upon such measurements. For example, the level of ovarian (female sex) hormone in the blood is measured by its effect upon the sex apparatus of an immature animal such as a mouse, rat, or rabbit.

Representative quantitative tests will be described briefly. No attempt will be made to give exact technical procedures. The tests can ordinarily be carried out only after a certain amount of technical training in the

procedures, and for discussion of these procedures the reader must be referred to textbooks on the subject.

1. **The basal metabolism test for thyroid function.** This test aims essentially at finding the rate at which the metabolic processes go on in the body. In order to understand the nature of the test, let us represent the metabolic processes as chemical changes in the nature of oxidation processes:



Since the rate of this process is primarily controlled by thyroxin from the thyroid, acting in the nature of a catalyst, as some have expressed it, thyroid function may be measured if the rate of this process can be measured. It is conceivable that the rate of liberation of waste products or utilization of fuel would be proportional to metabolism or to the rate at which the whole process goes on. But these are hard to measure. Because it is easily determined, oxygen consumption is usually measured. It is easily measured because it is consumed alike by everybody, because it is not stored to an appreciable extent in the body, and because it can be designated in easily measurable units. A test of metabolic rate is, therefore, carried out by measuring the amount of oxygen which a person consumes in a given time. In order that all tests may be taken under the same conditions and that there shall not be present varying factors which may influence the metabolic rate, the test is done in a resting state after a fast of at least six hours. Basal metabolism machines for making such a test consist of an arrangement for administering pure oxygen to the subject and of measuring the amount of this consumed over a given time, usually eight or ten minutes. The

amount consumed is often recorded graphically by the drop in a record line which is made on a chart attached to a revolving drum. One of these charts was presented in the discussion of metabolic changes in fatigue (see Fig. 25). The result of a metabolism test is usually expressed as a *basal metabolic rate*, which is the ratio of the amount of oxygen actually consumed in a unit time compared with the amount which a person should normally use according to his age, height, and weight. High metabolic rates are expressed as plus values over zero, as +25; low metabolic rates, as minus values below zero, as -21.

2. **Blood-sugar test for insulin production.** As has already been suggested, pancreas function in its production of the hormone insulin is measured in terms of blood-sugar content. Blood-sugar content can be taken as a fair indication of production of insulin, since rarely are there marked variations in blood sugar due to other causes. The blood-sugar test is a quantitative chemical test in which the number of milligrams of blood sugar per hundred cubic centimeters of blood is determined. In the normal individual, the blood contains about 100 mg. of sugar per hundred cc., limits of 90 and 120 mgs. being widely accepted as lower and upper normal limits. Excessive amounts of blood sugar are practically always indicative of insulin deficiency. Excessively low blood sugar is usually indicative of amounts of insulin above normal, usually met only in excessive artificial administration of insulin in diabetics. Blood-sugar level and insulin function afford excellent examples of the dependence of mental life upon chemical balance. In most cases of diabetes, the mental disturbances are usually of a mild type, perhaps only a mild depression. As the blood-sugar content grows higher and higher, coma, with complete

loss of consciousness, finally occurs. Inject into the individual some insulin, and his blood-sugar level decreases, the balance is swung in the other direction, and he resumes mental functioning. Continue to give insulin, and the individual will again pass into a state of coma due to too little blood sugar. The balance can be swung back to normal by administration of glucose (sugar). It seems from this that, in order to maintain normal mental functioning, a remarkable chemical balance is necessary with respect to insulin production and blood-sugar level.

3. Measurement of parathyroid function. Secretion of parathyrin by the parathyroid glands may be measured in terms of calcium content in the blood. This may be considered proportional to the parathyroid function, since one of the chief roles of the parathyroid is that of controlling calcium metabolism in the body. Like the blood-sugar test, the calcium test is one of quantitative determination of the amount contained in a unit quantity of blood. Normal calcium figures are from 9 to 11 mgs. per hundred cc. of blood. A decrease in this index suggests parathyroid deficiency, and, in the presence of other indications of parathyroid disturbance, would call for parathyroid therapy.

4. Measurement of ovarian hormone. The ovary, or female sex gland, secretes an important hormone termed oestrin, which has a marked influence on the sex cycle and on changes incident to pregnancy. It is this hormone which is responsible for production of the oestrous state ("heat" in animals). The measurement of oestrin content in the blood is of importance in many clinical conditions in which sex function is disturbed. Quantitative estimation of oestrin level in the blood depends upon its effect in producing an oestrous state in a female animal from which the ovaries have been removed. The

test is made by administering to a mouse, from which the ovaries have been removed 14 days previously, concentrated blood from the individual to be tested. Twenty-four hours and forty-eight hours after the injection has been given to the mouse, an examination is made to determine whether the injection has produced an œstrous state. Since the œstrous state is manifested by changes

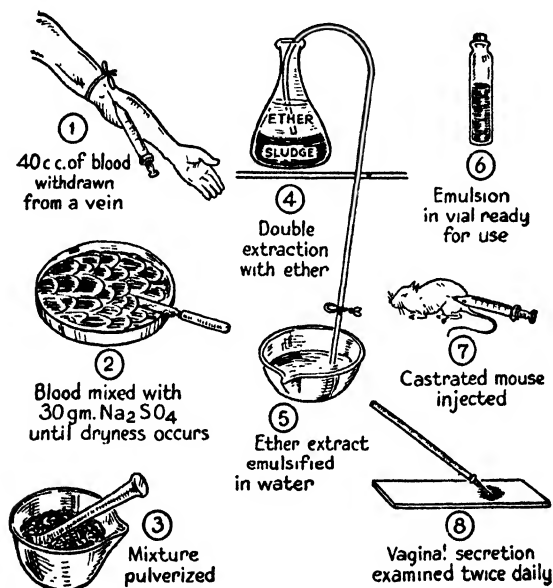


Fig. 28.—Technique of Blood-œstrin Determination.

taking place in the vagina, the test consists of an examination of vaginal smears. Mazer and Goldstein² have graphically depicted the technique of blood-œstrin determination as shown in Fig. 28. Blood-œstrin content at various stages in the sex cycle in the non-pregnant

² Mazer, Charles, and Goldstein, Leopold, *Clinical Endocrinology of the Female*, W. B. Saunders Co., Philadelphia, 1933, p. 156.

individual and in the pregnant individual is indicated in Fig. 29.⁸ Gonadal dysfunctions may show deviations from this normal curve. During pregnancy the œstrin level shows a continued rise, beginning early in pregnancy and reaching its peak just before birth of the child.

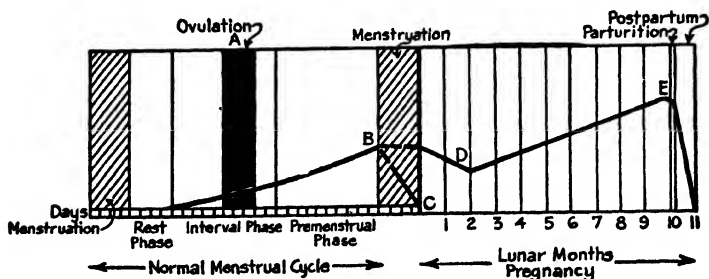


Fig. 29.—Blood-œstrin Content at Various Stages of the Sex Cycle.

A—œstrin occasionally demonstrable at this time in 40 cc. of blood.

B—œstrin demonstrable in 94% of normal women at this period in 40 cc. of blood.

C—œstrin content falls during menstruation.

D—œstrin content drops in first two months of pregnancy owing to excessive secretion.

E—œstrin demonstrable by injections of whole blood into animals.

5. Hormone tests of pregnancy. Hormone tests for pregnancy are generally based upon the presence in increased amounts, either in the blood or in the urine, of œstrin, which we have just discussed, or of anterior pituitary sex hormones. Since the anterior part of the pituitary gland secretes one or more hormones which have a part in the sex cycle and are essential to the normal functioning of the sex gland, the presence of these pituitary hormones or their variations are often just as indicative of changes in sex gland functions as are changes in

⁸ *Id.*, p. 166.

the sex hormones themselves. Their utilization in tests of pregnancy is possible because of their secretion in increased amount. During pregnancy, the blood level reaches such a high point that these hormones may also be thrown off in detectable amounts in the urine. Since the tests based upon urine analyses can be made as easily and with less inconvenience to the individual tested, they are the tests usually employed in the hormone determination of pregnancy.

If the presence of œstrin in increased amounts is used as a test for pregnancy, the procedure carried out is very similar to that described in the last section. The anterior pituitary sex hormone tests of pregnancy are carried out by injecting the urine of the suspected case of pregnancy into an immature female animal (mouse, rat, or rabbit). The ovaries and other parts of the reproductive apparatus of the animal are then examined, after a suitable length of time, to ascertain the effect of the injection. If the injection is from a pregnant individual and has contained anterior pituitary sex hormone, the immature animal will have been stimulated to a state of sexual maturity, the ovaries will show evidence of ovulation, and other changes indicative of sexual stimulation will appear in the uterus and uterine tubes.

These tests have been of immense benefit to the clinician in making possible very early diagnoses of pregnancy and in enabling distinctions to be made between pregnancy and other conditions which may simulate it.

CHAPTER XXVI

Physiological Measurements of Emotion

THERE are many reasons for desiring to measure the emotions or the changes indicative of emotional reactions. Because of the "emergency" nature of emotion, particular significance attaches to its relation to conduct. Emotion itself calls for immediate reaction with little deliberation and with little weighing of the alternatives in the situation, a procedure which is conducive to the most satisfactory, the most social, or the best reaction in the long run only when previous reactions have established a good precedent for similar situations. Our specific interest may be in the individual differences in emotional reaction—in the individual variations in degree of emotion resulting from the same stimulus or situation; or in the individual variations in range of stimuli capable of arousing emotion; or in the individual variations in the expression of the emotions. At another time, we may pursue the measurement of emotions as a means of discovering bases or causes for impulsive action, unsocial behavior, or abnormal conduct. Again, we may study the measurement of emotions as a means to increased knowledge of the nature of the emotion itself.

We have discussed in Chapter XXI the attempts, through verbal tests, to measure conduct involving emotional elements. These measurements are for the most part measures of conduct, of adjustment or maladjustment to one's surroundings. The conduct is often, per-

haps usually, related to the emotions, but the measurement is not a direct test of the emotional state. The measurements are not even in a medium applicable to the emotions themselves. The tests are verbal. Emotions themselves cannot be verbalized; they are a function of the unconsciously acting visceral, glandular, and autonomic systems of our make-up. What we need in studying many of the problems connected with emotions is a more direct measure of the changes which are characteristic of the emotional state.

I. The Nature of Physiological Measurements of Emotions

The measurements we are about to discuss concern the physiological changes which accompany or constitute the emotional state. A brief review of these would seem pertinent. In most emotional reactions, the body is prepared for an immediate and intense activity through the action of the autonomic nervous system. For the significance and biological usefulness of this response, we must go back in the evolutionary scale behind the human species. In the lower animal, the natural consequence of fear is flight; of rage or anger, combat. Both of these emotions lead to natural responses which make immediate demands upon the body for energy production. Nature has taken care of the situation by a mechanism of response, connected with the emotion, that prepares the body for the demands. In our own cases, we as a species may not always make the activity response—we may not always flee from the fear-producing stimulus or engage in combat with our rage-producing opponent; nevertheless, we have not lost the biologically developed tendencies to react in such a fashion. In strong emotions, our bodies are still prepared for the emergency responses.

426 Physiological Measurements of Emotion

Among these preparations, we find an increased activity in the circulatory system—a faster heart beat and a higher blood pressure—for more efficient transportation of fuel material and removal of waste products to and from the scene of activity. Blood is withdrawn from the digestive system to give the muscles a greater supply; hence digestive activity is in abeyance. Blood sugar is released from its storage places in the body to furnish extra fuel material. Sweat glands increase their activity to help in regulating body temperature in a situation likely to result in increased heat production. Breathing becomes more rapid to increase the oxygen supply for activity processes and for the removal of waste products in the form of carbon dioxide. Adrenalin is secreted from the adrenal glands to help in bringing about many of these changes.

We sometimes test such physico-chemical changes as these when measuring emotions. Before we describe in more detail some of the physiological tests, a few general considerations regarding the methods should be mentioned:

1. Practically all the physiological measurements of emotions necessitate the use of rather delicate types of instruments, many of which require a skill for their operation. This fact limits the usefulness and general applicability of the methods, and in the hands of unskilled workers makes for inaccuracy of results. In most of the work of measuring emotions through physiological changes, automatic methods of recording have been used. Such procedures are to be recommended wherever possible. The automatic recording is usually accomplished by the use of a *kymograph*—a drum revolving at a constant, known speed, on which is attached some kind of record paper. The changes, often amplified, are recorded

by a pen or writing point on the record paper. Such an apparatus is pictured in Fig. 30, which shows the recording of stomach contractions.

2. Many of the measurements can be made only under conditions of considerable inconvenience to the subject. Measurements of stomach contractions, for instance, usually necessitate the swallowing of a rubber bulb and tube.

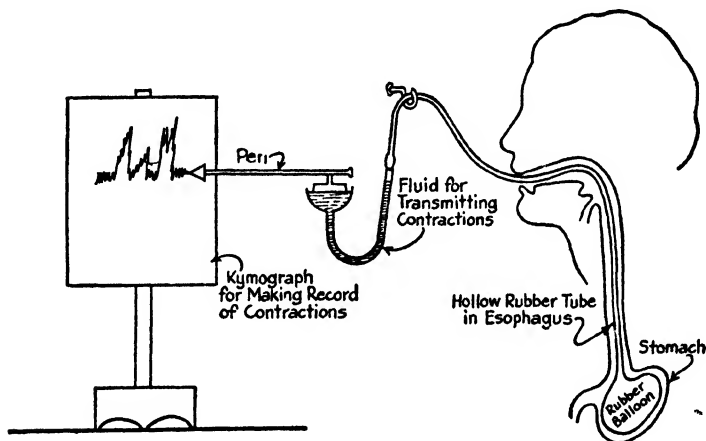


Fig. 30.—Diagram Showing Method Used to Record Stomach Contractions.

Biochemical measurements often necessitate the puncturing of veins to draw off a quantity of blood.

3. Most of the physiological measurements do not distinguish between the various emotions. They simply measure emotional excitement, and the records for fear may be indistinguishable from those for anger.

4. It is impossible to establish general standards of normality as useful norms by which to judge results of individual tests. Except for extreme or severe variations, a record during emotion means little except as compared with a "normal" taken on the same sub-

428 Physiological Measurements of Emotion

ject under conditions similar except for the emotional occurrence.

5. Emotional reaction is only one of many factors which may affect most of the physiological states measured. Blood pressure for instance is sensitive to many factors (listed in a subsequent paragraph of this chapter). The number of influences on a physiological state limit the usefulness of that state as a measurement of the influence of one factor, and make it useful at all only when the other factors are controlled.

With these facts in mind, let us briefly examine some of the physiological tests by which quantitative studies of the emotions have been made. We shall discuss (1) blood-pressure measurements; (2) breathing measurements; (3) measurements of digestive system activities; (4) psychogalvanic reflex measurement; (5) measurements of increased adrenalin secretion; and (6) measurements for detection of deception and lying.

II. Blood-Pressure Measurements of Emotion

Blood pressure is that pressure maintained against the walls of the arteries due to the force of the heartbeat, the resistance in the capillary blood vessels, and the pressure of the blood-vessel walls. This pressure, through changes brought about in the force and rate of the heart and in the character of the vessel walls, is sensitive to emotional excitement. Hence the utilization of blood pressure as a measure of emotion.

Blood pressure is ordinarily measured by noting the force or pressure necessary to collapse an arterial vessel and prevent the pulse from passing a given point. The brachial artery in the arm is the one commonly employed for measurement. The instrument used is known as a

sphygmomanometer. It consists essentially of (1) a wrapping band for the arm with a rubber bag inside which can be inflated and thus increase pressure on the arm; (2) a rubber bulb for inflating the bag in the arm band; and (3) a pressure gauge connected with the rubber bag. Measurement of systolic blood pressure (the measurement usually employed in studies of emotions) is made by inflating the bag to a pressure which is just barely sufficient to cut off the pulse that can be felt or heard below the point of application of the band. Such a pressure must be equal to the pressure of the blood, since any lower pressure in the band allows the blood to pass through and causes a pulse in the arm or wrist.

A number of investigations have been made of blood pressure in emotional states. All demonstrate that emotional excitement, short of a severe reaction of "shock," is accompanied by rise in pressure. Many of the studies have been aimed at applying the results to such practical problems as that of detecting lying. This utilization of measurements is discussed in a subsequent section of the chapter.

The drawbacks to blood-pressure measurements of emotions seem to be two: first, the fact that the measurements are not specific with respect to type of emotional experience; and second, the fact that blood pressure is affected by many things other than emotion, and that these cannot always be ruled out in the utilization of the measurement in studying emotions. Symonds,¹ in his discussion of blood pressure and emotions, lists 18 factors besides emotion that affect the pressure. All these should be borne in mind in the interpretation of emotional records. Symonds' list is given briefly here.

¹ Symonds, Percival, *Diagnosing Personality and Conduct*, The Century Co., New York, 1931.

430 Physiological Measurements of Emotion

1. Blood pressure varies directly with the volume of the blood.
2. Blood pressure varies directly with the energy of the heart.
3. Blood pressure varies with the elasticity of the blood vessels.
4. Blood pressure varies with the peripheral resistance.
5. Gravity causes variations in blood pressure.
6. Blood pressure is higher on the average in man than in woman.
7. Blood pressure varies with the size of the animal.
8. Blood pressure increases with age.
9. Breathing causes variations in blood pressure.
10. A rhythmical rise and fall of the blood pressure occurs, known as the Traube-Hering Waves.
11. Blood pressure is higher after food has been taken into the system.
12. Muscular exercise causes a rise in blood pressure.
13. Cold baths and hot baths produce a rise in blood pressure.
14. Blood pressure falls with severe fatigue.
15. Blood pressure is lower during menstruation and raised during pregnancy.
16. Pain causes a rise in blood pressure.
17. Cold produces a rise in blood pressure.
18. Certain glandular products cause marked changes in blood pressure.

From the above, it would seem obvious that conditions must be carefully controlled if presence of emotion or degree of emotion is to be detected by blood pressure changes. It also seems clear that we cannot compare blood pressure obtained at one sitting with pressure at another sitting, or pressure in one person with pressure in another. Even under the best of conditions, blood pressure will probably be of little indicative value except in rather intense emotions, in which the effect of the emotion would outweigh the lesser forces causing blood pressure to vary.

III. Measurement of Breathing

Breathing has several roles to play in its biological usefulness in the "emergency and preparatory" reaction of an emotional state. It is the means of intake of oxygen, which will be demanded in increased quantities in any emergency reaction of the organism. It is the means of exhalation of one of the most important end products in the elimination of waste material of activity. It is therefore logical for us to expect emotional excite-

TABLE XXVIII
EFFECTS ON RESPIRATION OF FEAR PRODUCED
BY FALLING

<i>Respiration Response</i>	<i>Informed Subjects</i>	<i>Uninformed Subjects</i>
Change in rate ..	20% decrease	12% decrease
Duration of change....	5 minutes	3 minutes
Change in I/E ..	281% increase	201% increase
Duration of change	5 minutes	3 minutes
Inspiration Stimulation	Marked	Marked

ment to be manifested by changes in breathing. The aspects of breathing usually studied are its rate, its depth, and the inspiration-expiration ratio (duration of inspiration divided by duration of expiration). Table XXVIII, from a study by Blatz,² gives the breathing changes for subjects stimulated to a fear reaction by the release of the chair in which they were sitting so that they were allowed to fall backward about 60 degrees. Some of the subjects were informed about the procedure; some were not. It will be noticed that rate of breathing decreased, but the I/E ratio increased, being over twice as great after excitement as before. These are common findings

² Blatz, W. E., "The Cardiac, Respiratory and Electrical Phenomena Involved in the Emotion of Fear," *Journal of Experimental Psychology*, Vol. VIII, 1925, pp. 109-132.

432 Physiological Measurements of Emotion

in emotions, although there have been variations in results reported by some investigators.

Breathing is subject to almost as great a variety of influences as blood pressure, and many are the same as those affecting blood pressure. Consequently, great care must be exercised in the control of conditions under which breathing changes in emotions are recorded.

IV. Measurement of Digestive-System Activity

Utilization of digestive-system activity in indicating emotional reaction goes, in story at least, back to the ancients. There is an old account that the ancients made those suspected of wrongdoing place dry rice in their mouths for a brief interval of time. If the rice came out dry, they were judged guilty. If the method is at all reliable, it is dependent upon emotional inhibition of salivary secretion in the guilty.

Cannon³ was one of the first actually to demonstrate experimentally that pleasant emotions are conducive to the best functioning of the digestive apparatus; and that agreeable surroundings, appetizing food, pleasant conversation, and a placid mind are all contributing factors to the normal flow of digestive juices and the normal muscular activity of peristalsis. He demonstrated, on the other hand, that pain, fear, rage, excitement, worry, and anxiety have antagonistic effects on digestive functions.

Cannon, and others following him, have employed three general methods for investigating digestive changes accompanying emotions. The first consists of the study of secretions of digestive juices through the observation in animals of the actions in a side pouch of the stomach

³ Cannon, W. B., *Bodily Changes in Pain, Hunger, Fear and Rage* (2nd ed.), Appleton-Century Co., New York, 1929.

which is brought to the outside by an operation. A few such studies have been made on human beings who, owing to disease of the esophagus, have had to have gastric fistulae for the introduction of food. Cannon found in dogs with esophageal fistulae and with side pouches opening to the exterior, that presentation of food caused flow of gastric juices even during sham feeding, but that after the dogs were aroused to fear or anger, there was no flow of gastric juices on presentation of food, even though the animals were hungry and ate with relish.

A second method utilized by Cannon enabled him to study peristaltic movements during emotional excitement. He fed his animals bismuth or some substance opaque to X-rays, and then observed the movements of their stomachs and intestines under various conditions. Thus he demonstrated the inhibitory effect of emotional excitement on peristaltic activity.

The third method, by the use of a rubber balloon which the subject swallows, records stomach contractions. The method is shown diagrammatically in Fig. 30.

These digestive measurements are presented for their theoretical interest rather than for their practical value as yardsticks in the detection and measurement of emotional reactions. The inconvenience to the subject which their use entails is sufficient to bar them from extended practical use.

V. Psychogalvanic Reflex Measurements

The term *psychogalvanic reflex* (or psychogalvanic skin reflex, since the changes are usually measured by electrodes applied to the skin) refers to certain electrical changes which accompany emotional excitement. These changes are usually recorded by applying two electrodes to the skin of the body at two points, as on two fingers,

434 Physiological Measurements of Emotion

or at two points on the arms, in such a way that the body forms a part of an electric circuit in which there is a sensitive galvanometer to record the changes in electric current. These changes were among the first physiological accompaniments of mental and emotional activity to be studied. Such a study was reported as early as 1879 by Vigouroux. Other early studies were conducted by Féré and Tarchanoff, their works being reported in 1888 and 1890 respectively. Since then the psychogalvanic reflex has been used in many investigations.

The main interest in the phenomenon as a test in the field of emotions has rested upon the fact that emotional changes often are paralleled by psychogalvanic reflex reactions. Many students of the problem have demonstrated these reactions, but there is some doubt as to the consistency of the results and the proper interpretation to be given them. A review by Paterson led him to believe that no characteristic psychological experience can be identified with any characteristic of the psychogalvanic reflex curve. Landis has summarized the whole matter quite capably, concluding that the reflex is one of a series of responses linked with the autonomic nervous system, but that it either may appear with no emotional accompaniment or may not appear when emotion is quite definitely present. He believes it worth studying as a phenomenon, but not as a test of emotion. Not all investigators are quite as pessimistic, however, as Paterson and Landis.

VI. Emotion Indicated by Measurement of Adrenalin in the Blood Stream

It is generally known today that emotional states cause a liberation of increased adrenalin from the ad-

renal glands. Why might we not utilize this physiological change as a measurement of emotion?

Cannon was the first to make such a measurement. In demonstrating the presence of adrenalin, he made use of the inhibitory effect of adrenalin on intestinal muscle contraction. In order to determine whether there was increased adrenalin accompanying emotional disturbance, cats were placed near barking dogs and were excited by them. Blood was taken from a vein of a cat before and after exposure to the dogs. Strips of intestinal muscle from a freshly killed animal had been arranged so as to contract rhythmically in a normal salt solution, each muscle having been attached to a recording device to show its contractions. Fig. 31⁴ shows what happened

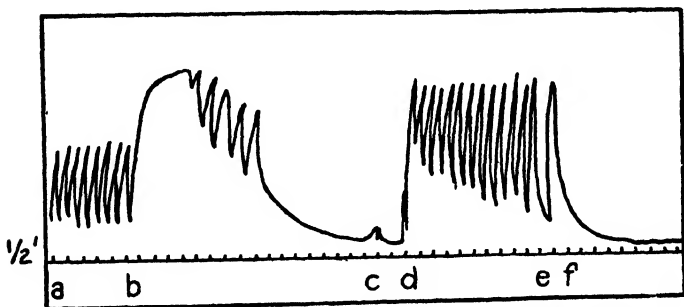


Fig. 31.—Record Indicating Inhibitory Effect of Adrenalin on Intestinal Muscle Contraction.

in the experiment. At (a) the solution in which the muscle had been beating was removed; at (b) blood from an excited cat was added; at (c) the "excited" blood was removed; at (d) blood from a quiet animal was added in its place; at (e) the "quiet" blood was removed; and at (f) "excited" blood was again applied. The contrac.

⁴*Id.*, p. 51.

486 Physiological Measurements of Emotion

tion records showed that blood from the vein after emotional excitement caused a relaxation of the muscle, but "quiet" blood caused little or no change in contraction. Cannon concluded from such experiments that the blood after emotional excitement contained more adrenalin, since no other substance to be found in the blood could have produced such an effect. In order to confirm the finding further, he removed the adrenal glands from one of his cats and found then that the blood, after the cat had been excited, produced no change in contraction of the test piece of intestine.

VII. Physiological Measurement of Emotion in Detection of Lying

Of the practical attempts to utilize physiological measurements of emotions, those which have caused most discussion have aimed to determine whether or not a person is lying in a given situation. From the scientific standpoint, the problem may seem a very small part of the study of physiology of emotions; but in the practical situations of criminal procedures there is often a real need for some device that will check the veracity of testimony and confessions. The instruments used in experimental work in this direction have earned the popular title of "lie detectors."

The theory underlying the work has been that when one is telling the truth there is no strong emotional reaction, and that any physiological changes of an emotional nature that are produced are relatively mild and lack marked fluctuations. On the other hand, when one is lying, or is deliberately concealing the truth, he is on his guard and is under an emotional strain, which is reflected in the physiological changes accompanying emotions and in the marked fluctuations in physiological

levels at points of crisis in the story or testimony. The obvious suggestion that the hardened criminal can lie without such changes is denied by the investigators.

Much of our information about physiological methods of detecting lying is derived from the investigations of Marston, Larson, and Landis. Marston studied in particular the possibilities of blood pressure. He claims that there are blood pressure curves typical of truth and lying. According to him, the curve of lying shows a steady rise as questions are answered and testimony is given, reaching a peak with a crisis in the testimony. Truth curves are smoother; the blood pressure records pursue a relatively level course with only slight and few rises. Marston claims that rises due to incidental factors in the truthful subject are always insignificant enough to cause little confusion. Larson has utilized in most of his work a combination of records, including blood pressure, pulse, and breathing. He claims exceptionally good results from utilizing Inspiration-Expiration Ratio in breathing (I/E). This ratio becomes larger, owing to relatively long inspiration, in the emotional strain of lying. Landis has experimented with both circulatory and inspiratory types of measures. Chappell in a more recent investigation has attempted to check some of the work of the earlier investigators. He points out that "deception curves" might be obtained under conditions other than those of deception, as when the subjects had been told that their intelligence was being measured.

Much of the experimental work has been done in the psychological laboratory with artificial, faked crimes. In a few instances, coöperation has been given by police authorities and actual criminal suspects have been studied. Marston, using systolic blood pressure, and Larson,

438 Physiological Measurements of Emotion

using a combination of blood pressure and breathing records, both claim over 90 per cent accuracy in detecting lying. Per cent accuracy of detection as reported by Landis is considerably lower, although he believes both blood pressure and I/E ratio possess definite diagnostic value for truth and lying. The best opinion seems to be that the extreme claims are too great to expect in practical applications. Undoubtedly much controlled investigation is needed before such methods can be of much use in actual criminal procedures. More experimentation is needed in real as contrasted with artificial laboratory situations; there needs to be some perfecting of instruments and technique; we must know more about the effects, on results, of familiarity of the subject with the procedure, of type of questioning, of type of surroundings in which questioning is done, and of personality make-up and emotional susceptibility of the subject. The method is not without considerable promise and, as an application of great practical significance, deserves these further studies.

Considerable discussion has recently centered about another interesting procedure which has been suggested for the detection of guilt. Reference is made to the interest in so-called "truth serums." These are essentially sedative drugs that, by forestalling the inhibitory action of the higher brain centers (which presumably direct the "lying"), leave the guilty one in an uninhibited state in which his responses to questioning will be the natural, truthful ones. The drug scopolamine has been utilized for this purpose; it is a drug often used in obstetrical anesthesia. It is probable that babblings during such anesthesia, of things ordinarily not told, led to the trial of the drug in detections of guilt. The following story,

with names omitted, depicts one of the more publicized cases in which the method was employed:

One morning ———'s men were called to a room in a city apartment hotel, where they found on a bed the naked corpse of a woman with cuts and bruises on her face, tooth marks on her body, blood on the pillow. Her clothing, torn in strips, was strewn on the floor. Identified as Mrs. F. H., she was found by the coroner to have died of a heavy blow on the head. She had been dead about twelve hours, but a woman's voice had answered the telephone in the room only two hours before the body was found. Perplexed, the policemen hunted the man with whom Mrs. H. had been living, a butcher named F. F. said he had not seen his mistress for two days, had spent the time in "night clubs." He had obviously been drunk. Asked if he would take scopolamine to refresh his memory, he agreed.

F. was taken to a hospital, and given four hypodermic injections of 1/150 grains each. For three hours he was incoherent. Then he revealed that he had downed several drinks with Mrs. H., tried to quiet her when she became boisterous, struck and choked her, tore her clothes. When she fell, struck her head against a metal bedpost, and lost consciousness, F. tried to revive her by biting her; he failed, remained several hours, then departed. What he did not tell while under the influence of scopolamine was that he had imitated a woman's voice when he answered the telephone. He was held for first-degree murder.⁵

Opinions range from very favorable ones which view such methods as capable, with extended skillful use, of eliminating much "third degree" brutality, to very adverse opinions which regard such procedures as unjust and unethical, and a defiance of human rights. These, of course, are ethical questions and have little bearing upon

⁵ *Time*, Vol. XXVI, No. 21, November 18, 1935, p. 55.

440 Physiological Measurements of Emotion

the scientific efficacy of such a method, which remains yet to be demonstrated thoroughly.

VIII. Summary

Throughout this chapter we have noted much which emphasizes that our knowledge of the physiology of the emotions is incomplete at present. The rapid progress that is being made today in the fields of endocrinology and biochemistry give us hope that the time is not far off when we shall know more about the factors we have been discussing. The rapidity of the progress will depend considerably upon the coöperation between psychologists and physiologists. The one group is inclined too often to talk glibly about the role of the endocrines and other physiological mechanisms in emotions without much definite knowledge as a basis; and the other group is often just as careless in dismissing without careful investigation the effects and physiological manifestations of what he terms psychic factors. Progress, we repeat, will depend upon careful study in this common ground of psychology and physiology. And with this progress we may expect the quantitative estimate of emotional states through their physiological manifestations to become more accurate and more practical.

CHAPTER XXVII

Tests in the Motor and Sensory Fields

MOTOR and sensory tests of the type to be discussed in this chapter have an importance to psychology for several reasons. First of all, they possess great interest from a historical standpoint. In one of the first chapters of this book, we noted the use of sensori-motor tests in the early investigations of individual differences. Many of these earlier tests have been admirably summarized and directions for their application given by Whipple in his *Manual of Mental and Physical Tests*, published in 1914. Motor and sensory tests have long interested the psychologist as a means of studying the relation between physical and mental characteristics. Motor and sensory deficiencies markedly affect social and emotional adjustments, achievement in school, and proficiency in work. Hence, the importance that attaches to them in the problems of clinical testing, of vocational and educational guidance.

The tests described in this chapter are selected from a large number of available motor and sensory tests. They have been selected because they are representative of the field, because they measure rather broad aspects of motor or sensory efficiency, or because they have been of particular interest or value to the psychologist.¹ In nature and

¹ For more extensive discussions of such measurements, the reader is referred to such works as Brace's *Measuring Motor Ability*, Paterson's *Physique and Intellect*, and Whipple's *Manual of Mental and Physical Tests*.

442 Tests in the Motor and Sensory Fields

general purposes, they overlap some of the purely physiological measurements described in Chapters XXIII and XXVI. The reader may in some instances wish to consider all these tests together.

1. **Cephalic index.** This is an index of head shape, calculated from measurements of head width and length made with a pair of head calipers (see Fig. 32):

$$\text{Cephalic Index} = \frac{100 \times \text{Width}}{\text{Length}}$$

It has been customary to designate as *dolichocephalic* the "long-headed" with an index of below 75; as *mesocephalic* the "medium-headed" with an index of 75 to 80; and as *brachycephalic* the "broad-headed" with an index of above 80.

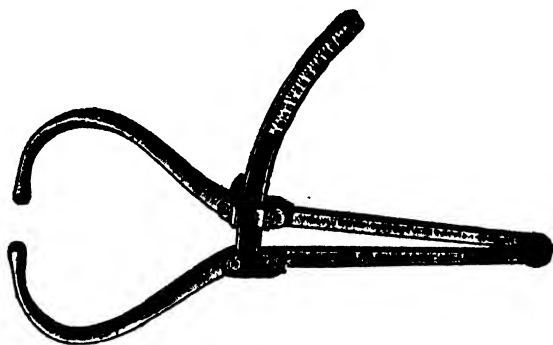


Fig. 32.—Head Calipers.

The cephalic index has interested chiefly the anthropologist in his search for measurements differentiating racial groups, and the psychologist in his search for physical measurements reflecting mental ability. In the early sensori-motor stage of ability testing, head measurements were among the favorite physical tests. Before the development of his intelligence scale Binet investigated the

possibilities of cephalic measurements as means of indicating intelligence quantitatively; in fact, the greater part of Binet's experimental output during the years 1901 and 1902 deals with cephalic measurement. His conclusions are based primarily upon measurements of about 250 children of ages 11 to 13. His subjects were selected on the basis of estimates of their ability made by their instructors, and were estimated as decidedly superior or inferior. His average head measurements favored, so far as head size is concerned, the intelligent group as compared with the unintelligent, but the differences were small and the overlapping great. Binet himself admitted at the time of his study that his head measurements could not be regarded as safe tests of intelligence for purposes of individual selection, and that except for the very extreme measures, head size and shape did not constitute satisfactory indicators of intellectual capacity.

Another of the early studies, similar to Binet's, is that of Pearson. He correlated head measurements with teachers' estimates of intelligence for over 5,000 subjects, including a group of college students, a group of twelve-year old boys, and a group of twelve-year old girls. The correlations for these groups were, respectively, $-.06$, $-.04$, and $+.07$.

More recent studies of this same relationship, made after the advent of intelligence tests, substantiate the results of the earlier studies in which intelligence was only estimated. Most of the reported correlations between cephalic index and intelligence test measures are below $.10$.

2. The strength tests. Various strength tests have been used in psychological and physiological studies. The studies have had widely different purposes. Some

444 Tests in the Motor and Sensory Fields

have aimed at evaluating athletic and gymnastic ability; some have depicted growth and maturation in physical capabilities; some have aimed at measuring physical traits important in industrial and vocational performances; and many conducted in the psychological laboratory have aimed at finding the relationships between strength and other traits and abilities.

One of the most popular of the strength measurements in the psychological studies is that of measurement of

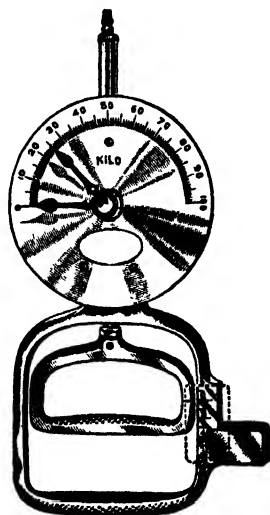


Fig. 33.—Dynamometer for Measuring Hand Strength.

hand strength, made by use of the hand dynamometer. (See Fig. 33.) The subject grips the two handles in such a way that the second phalanges of the fingers press against the inner handle. The strength of the grip is indicated in kilograms on the dial. Norms for strength of grip on such an instrument are given in Table XXIX.

A number of investigations have been made of the relationship between strength of grip and mental ability.

With children the correlations reported have usually been positive. Johnson ² reports a correlation as high as .71 between strength of grip and mental age for 262 children between the ages of 3 and 13. However, with older groups, and with the younger groups when age is held constant, the correlation is near zero.

TABLE XXIX
NORMS FOR STRENGTH OF GRIP (IN KILOGRAMS) ³

AGE	Boys		Girls	
	Right Hand	Left Hand	Right Hand	Left Hand
6	9.21	8.48	8.36	7.74
7	10.74	10.11	9.88	9.24
8	12.41	11.67	11.16	10.48
9	14.34	13.47	12.77	11.97
10	16.52	15.59	14.65	13.72
11	18.85	17.72	16.54	15.52
12	21.24	19.71	18.92	17.78
13	24.44	22.51	21.84	20.39
14	28.42	26.22	24.79	22.92..
15	33.39	30.88	27.00	24.92
16	39.37	36.39	28.70	26.56
17	44.74	40.96	29.56	27.43
18	49.28	45.01	29.75	27.66

The dynamometer has been used in several studies of fatigue. Fatigue "curves" have been constructed from successive dynamometer records, and their characteristics studied. In the investigation of fatigue recently sponsored by the Society of Automotive Engineers ⁴ the hand dynamometer was studied as a possible means of measuring degrees of fatigue. It was finally discarded

² Johnson, B. J., *Mental Growth of Children in Relation to Rate of Growth of Bodily Development*, E. P. Dutton and Co., New York, 1925.

³ Whipple, G. M., *Manual of Mental and Physical Tests, Simpler Processes*, Warwick and York, Baltimore, 1914.

⁴ *Journal of Society of Automotive Engineers*, Vol. XXVI, No. 4, April 1930.

446 Tests in the Motor and Sensory Fields

in these studies, since strength of grip showed no definite variation with relatively mild degrees of fatigue.

In a recent study⁵ of handedness the hand dynamometer was utilized as one basis for determination of degree of handedness—the degree of handedness being taken as the ratio of performance by the right as compared with the left hand.

3. Tests for sensory acuity. Sensory efficiency, especially in hearing and vision, is of such great importance in our daily lives that it is only natural that standard instruments should have been developed to measure it. Measurements are often made by the ophthalmologist or the otologist, but from a very early period psychologists have been interested in sensory studies.

In the psychological studies sensory capacity is commonly measured by two general methods. By the first we find the least stimulus which can be perceived by the subject; this minimum constitutes the *absolute threshold* or *limen of sensitivity*. The dimmest light that can be perceived constitutes the absolute threshold of vision; the faintest sound that can be heard, the absolute threshold of hearing. By the second method we find the smallest difference between two stimuli that can be distinguished by the subject. This difference constitutes the *just noticeable difference* or the *differential threshold*. To illustrate, if we are investigating pitch sensitivity in hearing tones, we determine the smallest difference in pitch that the subject can distinguish between two tones. Absolute threshold and differential threshold are not perfectly correlated. We do not always find low differential thresholds in those with low absolute thresholds. Absolute threshold may be said to be a measure of gen-

⁵ Roos, Mary M., Unpublished Thesis, The George Washington University.

eral sensory efficiency; differential threshold is particularly related to matters of fine sensory appreciation. Seashore has stressed such measurements in hearing in his studies of musical talent.

Auditory acuity is commonly measured approximately by "voice tests," "watch tests," or "tuning fork tests," in which records are made of the distances at which the various sounds can be heard. These methods are obviously hard to standardize, and the results are, therefore, likely to be rather inaccurate. More accurate measurements of auditory acuity make use of audiometers, instruments designed for accurate production of sounds of varying loudness or intensity. One of the best of these is the audiometer designed by Seashore. The stimuli consist of series of clicks of graduated intensity, which are heard through telephone receivers. Space does not permit here a detailed account of this instrument.

Visual acuity is most frequently measured by a standard set of test cards from which the subject reads letters, numbers, or designs at a given distance. The most familiar device is the Snellen Chart. Rows of letters varying in size are printed on the chart and are to be read by the subject. Given on the chart is the distance at which the normal eye should be able to read the letters of each row. Visual acuity is expressed as a fraction, the numerator of which is the distance the subject stands from the chart (usually 20 feet), and the denominator of which is the distance value of the smallest letters that can be read. For example, if a person can read at 20 feet the letters normally readable only at 15 feet his vision is 20/15 (better than normal); if he can read at 20 feet no smaller letters than can normally be read at 30 feet, his vision is 20/30 (poorer than normal).

Other standard visual acuity tests include the Ewing

448 Tests in the Motor and Sensory Fields

Test, a series of charts designed to avoid some of the limitations of the Snellen Test; and the McCallie Tests, a series of charts designed especially for testing children and illiterates.

4. Color blindness. Color blindness is an interesting defect in the sensory field that has been the subject of a number of psychological studies. There are various forms of color blindness, the commonest of which is red-green blindness, in which these two colors are confused with each other and with other colors of like brightness and saturation. Rarer forms of color blindness are the yellow-blue and the total forms.

Two standard tests for color blindness are the Holmgren Woolens Test and the Ishihara Test. In the former the subject must pick out from numerous small skeins of wool those which resemble in color three large skeins which are green, rose, and red in color. Errors in picking out resembling colors indicate color blindness. The Ishihara Test consists of sixteen color plates designed so that the normal person can easily distinguish numbers on certain of the plates, whereas the color-blind person has difficulty in distinguishing the numbers or fails to see them altogether.

5. Steadiness tests. Steadiness tests have been popular in the psychological laboratory for many years. They have been employed in studies of individual differences and in studies of age and sex differences. They have been used in studying the effects of fatigue and of drugs such as alcohol, tobacco, and caffeine.

One of the oldest and most commonly used of these tests is the hand steadiness test designed to measure muscular control and control of tremor. One design of this test worked out and standardized by Swope of Purdue University and used in a series of automobile driving

tests is illustrated in Fig. 34.⁶ The subject takes the test by inserting the stylus successively in each hole, beginning with the largest and moving forward at the rate of about one hole per second. The stylus is pushed to a back stop at the depth of an inch, the subject being careful not to touch the brass plate with the stylus. Should the edge of the hole be touched with the stylus,



Fig. 34.—Instrument for Testing Hand Steadiness. The instrument is hung in a rigid position about 36 in. from the floor of the car. Its position requires that the operator sit forward on the seat and thus have no support for his back.

a buzzer inside the cabinet rings, and the number below the hole where contact is made is recorded against the subject's score.

The Moss *Wobblemeter* is a recently designed instru-

⁶ Reproduced by courtesy of Ammon Swope and the *Journal of the Society of Automotive Engineers*.



Fig. 35.—Moss Wabblemeter for Measuring General Bodily Steadiness. The platform on which the subject stands moves in a vertical direction upon a central pivotal point as the weight is shifted from one foot to the other and from heel to toe. This motion is totalized and registered on two counters. The count increases with fatigue.

ment for measuring general bodily steadiness. The instrument was designed in connection with a research study of fatigue produced by automobile driving, sponsored by the Society of Automotive Engineers, and was utilized as a means of indicating degree of fatigue. Fig. 35 indicates the nature of the Wabblemeter. The platform on which the subject stands moves upon a central pivotal point as the weight is shifted from one foot to the other and from heel to toe. This motion is totalized in the instrument and registered on two counters.

Moss has reported a study of diurnal variations in steadiness as measured by this machine, and a number of studies of effect of fatigue produced by automobile and airplane riding and driving upon steadiness.⁷ The diurnal variations are shown in Fig. 36. There is a fairly general opinion that people are much steadier in the morning than they are in the middle of the afternoon; this seems to be far from the truth in this study. As a matter of fact, the most unsteady records are those of the early morning hours. The eight o'clock morning records were on the average about 60 per cent higher than the records at two in the afternoon. The increase in steadiness over part of the day probably has much of the nature of a "warming-up" period, such as the baseball player might find necessary. Such a study of diurnal variations is of importance in interpreting fatigue records or records taken for other purposes throughout a day. Moss's studies of fatigue from travel in various types of vehicles show progressive decreases in steadiness, the actual amount of decrease being related to type of

⁷ See *Journal of the Society of Automotive Engineers*, Vol. XXVIII, No. 5, May 1931.

452 Tests in the Motor and Sensory Fields

vehicle, speed of traveling, type of road, and weather conditions.

6. Reaction time. Reaction time is a measure indicating the speed with which we respond to stimuli, the term ordinarily being applied in those situations calling for immediate responses. This speed is dependent pri-

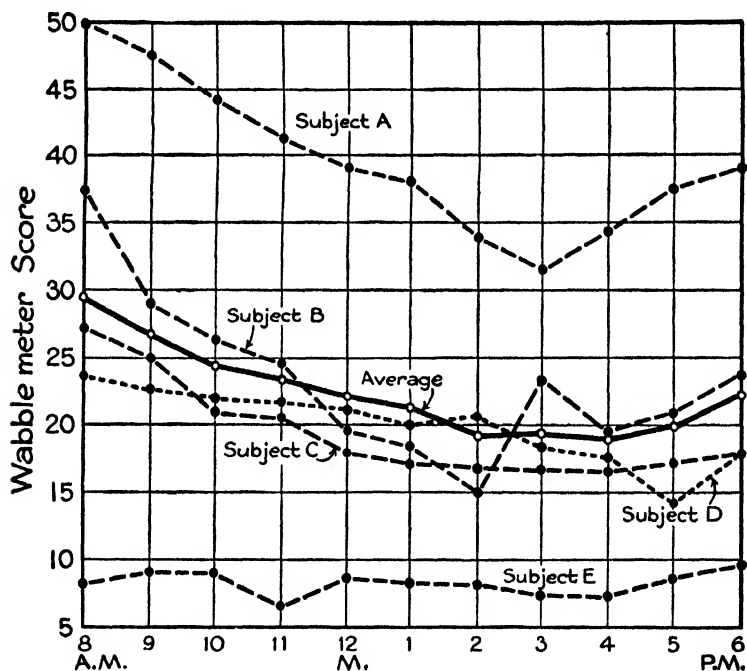


Fig. 36.—Diurnal Variations in Steadiness.

marily upon neurological factors of our makeup—the speed with which stimuli affect sensory nerve endings, the speed with which neural impulses are conducted, and the speed with which synaptic connections are made. Reaction times vary with the stimuli and with the responses to be made. They may be very short (only a few

thousandths of a second) for simple motor responses to single, elementary stimuli, as a blink of the eye to a visual stimulus. They may be considerably longer where a more complex response is to be made; or where a choice of responses is involved, as in the tapping of one key if a red light is given and another key if the stimulus light is green. The reaction time for the same response to the same stimulus is not the same for all individuals. In fact, marked individual differences occur in reaction times.

Theoretical interest in reaction time measurements has been in evidence for a long time. Studies in this field were favorites in early psychological laboratories and in the early psychological studies of individual differences. Neurologists also concerned themselves with studies of reaction time, particularly in its relationship to neural conduction. Studies along this line were reported by Helmholtz as early as 1850.

With the growth of applied psychology, emphasis has been placed upon reaction time as an important element in many practical, everyday situations. The ability to drive an automobile safely, to react promptly in the emergency situations of driving, is directly related to one's speed of reaction. In industry we find some men attempting to operate machines that require faster reactions than they are able to make, while others are operating machines that do not demand their best in speed of reaction. This lack of agreement between the demands of the job and the ability of the man doing the job is an important cause of fatigue and boredom felt by workers. The individual learning to operate a typewriter must possess a certain minimum reaction time in motor response and in eye-hand coördination. All these examples emphasize the importance of reaction time in practical affairs and should point out to us the value of reaction

454 Tests in the Motor and Sensory Fields

time measurements in relation to accident prevention, placement in industry, and vocational guidance.

During the past few years there have been several attempts to develop a reaction time instrument which would be of especial value in testing reaction time as it is involved in operating an automobile. Most of the attempts have aimed at developing a means of measuring the time it takes one to make the response of putting on the brake at a given signal. Since these instruments possess a greater practical value than many of the earlier instruments, we shall examine one or two of them as examples of measurement in the field of reaction time.

One of the earliest means of measuring the reaction time involved in putting on the automobile brake at a given signal was worked out by Moss and Allen. Their measurements were made by having the individual move his foot from the accelerator to the brake at an auditory stimulus. The stimulus was produced by the firing of a gun attached beneath the running board of the car in such a fashion that at the moment the gun was fired a mark of red lead was made on the road. When the driver pressed the brake pedal, more red lead was released and made another mark on the road. By maintaining a constant speed in the car and measuring the distance between the two red marks on the road, the time elapsing between the signal and the brake application could be calculated.

More recently, in connection with experimental work carried on for the Society of Automotive Engineers, Moss and Brown have developed another instrument for measuring reaction time. This instrument simulates the automobile situation in that the reaction to be made is the moving of the foot from one pedal representing the accelerator to another pedal representing the brake pedal,

although the actual test is carried on in the laboratory. Reaction is made to a visual stimulus produced by a light in a bulb in front of the subject. This instrument is illustrated in Fig. 37.



Fig. 37.—Moss-Brown Reaction Time Instrument.

456 Tests in the Motor and Sensory Fields

Rackley⁸ used this instrument in a series of studies of reaction time as related to various other factors. Some of his findings are fairly typical of results which have been obtained in various reaction time studies.

He studied the relation of reaction time to mental ability, using a group of university students, a group of high school students, and a group of inmates of the National Training School for Boys in Washington, D. C. The first two groups were measured in mental ability by the use of group intelligence tests; the third group was tested by the Stanford Binet Scale. The correlations between intelligence test records and reaction times are given in Table XXX. It will be noticed that with the

TABLE XXX
CORRELATIONS BETWEEN INTELLIGENCE AND
REACTION TIME

<i>Group</i>	<i>Correlation</i>	<i>P. E.</i>
University	+ .13	.03
High School	+ .23	.06
Training School	+ .55	.05

university students the correlation is lowest, being close to zero; while with the Training School group there is a significant positive correlation. The difference in relationship in the three groups is probably related to the homogeneity or heterogeneity of the group. The group showing the lowest correlation is the homogeneous, that is, the most alike in intelligence; whereas the group showing the highest correlation is the most heterogeneous, or widely different in intelligence. The National Training School group ranged from an I.Q. of 39 to an I.Q. of 113. The university group ranged from 90 to 125 I.Q., a difference in spread of the two groups of 39 points. From

⁸ Rackley, Lloyd Ernest, *An Experimental Study of the Factors of Reaction Time as Exhibited by a Cross-section of the Population of Washington, D. C.*, unpublished thesis, The George Washington University.

such a study we may conclude that in an unselected population there is a moderate positive correlation between reaction time and intelligence. In a selected group, particularly one selected from the upper levels of intelligence, there is practically no relationship between reaction time and intelligence.

The effect of practice on reaction time was studied on 147 subjects altogether. One group of these subjects took 15 trials each, another group took 30 trials each, and a third group took 20 trials a day for 20 days. Rackley concludes from his study of these groups that with practice there is a small but definite reduction in reaction time. From this conclusion we might infer that long reaction times can be shortened somewhat by attention directed specifically toward practicing the act involved. The amount of reduction, however, is too small to indicate that such practice procedures would be of much avail in the practical situations in which an individual's reaction time may be too long.

Among the other results of Rackley's investigations, with a brief conclusion about each, are the following: A slight difference in the reaction time of *white and negro* subjects was found, the difference being in favor of the whites. No significant difference in reaction time of *experienced and inexperienced* automobile drivers was found. An initial reduction in reaction time following *severe physical exercise* was found. A small but significant difference in the reaction time of the *two sexes* was indicated, in favor of men. Some variation in average reaction time at *different hours of the day* was found, longer reaction times tending to occur very early in the morning as compared with trials taken later in the day. It is interesting to note that this conclusion agrees with the findings which we quoted from the studies of steadiness.

INDEXES

Subject Index

A

Accident prevention, 306-309, 393, 454
Accomplishment quotient, 223
Achievement:
 measurement, 4, 40, 207-239
 college, 217-219
 elementary school, 213-214
 high school, 214-217
 professional schools, 227-239
 tests, 207-239
 comprehensive, 220-222
 examples, 213-223
 nature, 207, 223-226
 psychology's contribution to, 208-213
Activity interests, 200
Adaptability tests, 290
Adaptation board, 75
Adjustment questionnaire, 358
Adjustment tests, 358
Admission to college, tests in, 120-121
Adrenalin, 434-436
Adults, measurement of intelligence in, 107
Age scales, 65-72
American Council Psychological Examination, 107
American Psychological Association, 35
Annoyance test, 365-367
Aptitude tests, 15, 40, 143-204
 engineering, 173
 law, 173
 medical schools, 174-186
 professions, 172-186
 teaching, 173-174
Army Alpha Test, 18, 36-37, 42-49, 80-81, 105
Army Beta Test, 36-37, 50-56, 77
Army Performance Test, 37, 76
Army testing, 37, 265, 268
Arteriosclerosis, 409
Art tests, 153-161
Ascendance-Submission Scale, 330

A-S Reaction Study, 330
Assembly tests, 167, 168
Association of American Medical Colleges, 175, 185
Association tests, 115-117, 200-201, 368
Attitude tests, 388-389
Automobile driving, tests for, 448-454

B

Bacteriology test, 229-239
Basal metabolism, 400, 401, 408, 418
Behavior:
 control, 10
 prediction, 10
 tests, 373, 380-386
 understanding, 10
Binet Scale (*see* Binet Test)
Binet-Simon Test (*see* Binet Test)
Binet Test, 18, 24, 27-33, 36, 65-72, 110-113, 203-204
Blood cell measurements, 400, 405
Blood pressure, 388, 398, 409, 428-430
Blood sugar, 399, 406, 419
Bookkeeping tests, 250
Boston Elevated Railway, tests used in, 306
Boston Psychopathic Hospital Memory Test, 114
Brachycephalic, 442
Breathing, measurement of, in emotions, 431-432, 438
Bricklaying, motion study in, 310

C

Carbon-dioxide combining power of blood, 399, 406
Casulist board, 74
Cephalic index, 442-443
Character education, 373, 386
Character Education Inquiry, 375, 379

Cheating tests, 381-386
 Chemistry tests, 217-219, 251
 Civil Service Commission, tests of, 107, 132, 266-271, 286
 Classification questions, 213
 Classification tests, 220-222
 Clerical tests, 290
 Coefficient of correlation, 21
 College Entrance Examination Board, examinations of, 224
 College grades, relation to intelligence tests, 122-123
 College intelligence tests, 104
 College of Physicians and Surgeons, objective tests in, 229
 Colloidal gold test, 407-408
 Color blindness, 448
 Completion questions, 212
 Composition scale, Willing's, 220
 Comprehensive achievement tests, 220-222
 Conduct knowledge tests, 373, 375-380
 examples, 376-379
 value, 379
 Consonance, measurement of, 145, 147
 Constancy of mental ability, 137-138
 Correlation, 21
 Cretin, 62
 Criminology, tests in, 133-135, 371, 372
 Criteria, test, 21, 284-285
 Critical scores, 279
 Cube test, 75, 76

D

Dearborn Tests, 79
 Deception tests, 381-386, 436-440
 Deficiency, measurement of, 59-81
 Delinquents, use of intelligence tests with, 133-135
 Dementia praecox, 112, 117, 363, 409
 Department store management, tests in, 311-317
 Descriptive rating scale, 327-328
 Design tests, 77-79
 Deterioration, intellectual, 110, 112
 Diabetes, 406, 419
 Diagonal test, 75

Digestive-system changes in emotions, 432-433
 Digit-symbol test, 76
 Disciplinary cases, use of tests in, 120
 Dishonesty, tests of, 381-386
 Dislikes, 187-189
 Dolichocephalic, 442
 Dynamometer, 444

E

Educational age, 222
 Educational quotient, 222
 Efficiency:
 measurement, 240-253
 ratings, 241
Elemente der Psychophysic, 26
 Emotional adjustment, measurement of, 358-362
 Emotionality score, 369
 Emotional maturity scale, 330
 Emotional reactions, measurement of, 4
 Emotional tests:
 physiological, 424-440
 verbal, 357-371
 Employees:
 selection, 124-133
 tests in control, 303-318
 Employment tests, 273-302
 Endocrine tests, 410-423
 Energy, 185
 Entrance credits, medical school, 184
 Entrance tests, college, 120-121
 Epilepsy, 112
 Ergometer, bicycle, 397
 Ethical discrimination test, 368, 375
 Experimental psychology, 25
Experimental Study of Intelligence, 30
 Extroversion, measurement of, 362-365

F

Fatigue:
 in industry, 262
 measurement, 393-402
 Feature profile, 75-76

Feeble-mindedness (*see* Mental deficiency)

Form board test, 72

G

Genius (*see* Superiority)

George Washington University Series tests, 105, 214, 217, 228

Glands:

- adrenal, 411
- gonads, 411
- importance in psychology, 410
- pancreas, 412
- parathyroid, 412
- pituitary, 411
- sex, 411
- thyroid, 411

Glandular function tests, 410-423

Grades, relation to intelligence tests, 122-124

Graphic rating scale, 328-329

Group tests, intelligence, 32-34, 41, 80-81

H

Haggerty Intelligence Tests, 39, 79

Haggerty Reading Examination, 213-214

Handedness, 446

Handwriting scale:

- Ayres, 220-221
- Thorndike, 220

Head measurements, 442-443

Healy Puzzle Test, 75

Hemoglobin tests, 405

Hereditary Genius, 80, 138

Heredity, 138-139

History of measurement, 25-41

Background Period, 25-28

Binet Period, 28-32

Maturing Period, 39-40

Post-War Period, 32-34

Preliminary Period, 32-34

Present Period, 41

World War Period, 34-37

Holmgren Woolens Test, 448

Honesty tests, 332, 381-386

Hormone tests, 418-422

Humor, test for, 344

Hydrocephalic, 62

I

Identification questions, 212

Idiosyncrasy score, in emotions, 369

Idiot, 61

Imbecile, 61

Individual tests, 65-72

Industry, measurement in, 257-268

during the World War, 265-266

early experiments, 262-265

foundations, 257-262

testing of applicants, 266-271

Information tests:

interest in agricultural engineering, 196

interests, 196-199

mechanical interest, 197

play interests, 199

social interests, 197

vocational interests, 198

Insane, tests for, 109-117, 373, 403-409

Inspiration-expiration ratio, 431, 438

Insulin, test for, 419

Intelligence:

constancy, 137-138

distribution, 92, 94

measurement:

in deficiency, 59-81

in superiority, 82-108

quotient, 19, 69-70

tests, 4, 15, 23-41, 59-139, 268-269

college, 104-107, 119-123

for adults, 107-108

for deficiency, 65-81

group, 32-34

industry, 268

of the insane, 109-113

relation to college grades, 122-123

research, 136-139

schools, 95-104, 123-124

specialized, 269

uses, 118-139

vocational selection, 124-133

with delinquents and criminals, 133-135

Intensity, measurement of, 145-146

Interest inventories, 189-195

examples, 189

scoring, 193

Interest inventories (*Cont.*):

selection of items for, 193
 Strong's Inventory, 189
 validity, 194

Interests, 187-204

association tests for measuring,
 200
 children's, 195
 information tests for measuring,
 196
 inventories for measuring, 189
 objective, 188
 questionnaires for measuring,
 189
 rating scales for measuring, 195
 subjective, 187
 vocational, 189-196

International Intelligence Test, 78,
 97-101

Interne ratings, 182-183

Interviews for medical students,
 183, 184Introversion, measurement of, 362-
 365Inventories, interest (*see* Interest
 inventories)

Involution melancholia, 363, 414

Iowa High School Content Exam-
 ination, 222

Ishihara Test, 448

J

Job analysis, 275-276, 282

Job efficiency, measurement of,
 240-253

psychological tests, 250-253

service ratings, 241-250

K

Kahn Test, 404

Kent-Rosanoff Association Test,
 115-117

Kymograph, 426-427

L

Laboratory, psychology, 27

Laboratory tests in mental disor-
 ders, 403-409

Lange's test, 407-408

Lie detector, 436

Likes, 187-189

Limen of sensitivity, 446

Literature test, high school, 214-
 217

Lying, tests for, 381-386, 436-440

M

MacQuarrie Test for Mechanical
 Ability, 169, 171Macy Department Store tests, 311-
 317

Mania, 363

Manic depressive psychosis, 117

Manikin test, 75, 76

Man-to-man rating scale, 325-326

Mare and Foal Board, 73

Marietta Apparatus Company, 169

Matching test, 212

Maze tests, 76-79

McAdory Art Test, 157-161

McCallie Tests, 448

Measurement:

character, 372-380

development, 6

emotions, 357-371, 424-440

fatigue, 393-402

history, 23-56

in control of employees, 303-318

in everyday life, 5

in higher education, 228

in motor and sensory fields, 441-
 454

in physical sciences, 13

in psychology, 10

in selection of employees, 280-
 302

interests, 187-204

morality, 372-389

origin, 6

personality, 321-389

problems, 11

reaction time, 452-454

social attributes, 334-356

steadiness, 448-452

teaching efficiency, 251-253

value, 3-12

what is, 3

Mechanical aptitude:

measurement, 162-171

tests, 15, 40, 167, 169, 171, 197

Mechanical engineering tests, 228

Mechanical interest tests, 197

Medical aptitude test, 174-186

development, 174-175

nature, 175-180

relation to interne ratings, 182-
 183

Medical aptitude test (*Cont.*)
 relation to medical school grades,
 180-182
 sample questions, 176-180
 validity, 180-186
Medical Examinations, State
 Board, 175
Medical school:
 admission to, 174-186
 grades, 180-182
 prediction of success and failure
 in, 181-184
Meier-Seashore Art Judgment
 Test, 154-157
Memory:
 for names and faces, 335, 342
 tests, 114-115
Mental age, 32, 59-60, 67, 69
Mental Alertness Test, 105-106
Mental deficiency:
 among criminals and delinquents,
 134-135
 definition, 59-60
 examples, 71-72
 measurement, 59-81
 problems, 62-63
 tests, 63-81
Mental disease, glands in, 414-416
Mental disorders, tests in, 403-
409
Mental growth, 136-137
Mental Hygiene Inventory, 35-39
Mental Hygiene Test, Colgate, 359
Mental tests (*see* Intelligence
 tests)
Mesocephalic, 442
Metabolism, 400, 401, 408, 418
Microcephalic, 62
Minnesota Assembly Test, 169
Minnesota Mechanical Ability
 Tests, 167-169
Minnesota Paper Form Board,
 168-169
Moron, 61
Motility, 321
Motion study, 309-311
Motormen, tests for, 263-265, 306
Motor tests, 441-454
Multiple-choice questions, 211
Musical aptitude:
 effect of training on, 153
 inheritance, 153
 measurement, 144-153
 racial studies, 153
Musical memory, measurement of,
 145, 147

N

National Intelligence Test, 39, 81,
 95-98
Neurasthenia, 359, 409
Norms, 279
Numerical rating scale, 326-327

O

Occupations, intelligence limits for,
 128-130
Oestrous state, 420
Opinions tests, 386, 388-389
O'Rourke Mechanical Aptitude
 Test, 170-171, 197
Otis Intelligence Tests, 38, 101-
 104, 169
Ovarian hormone test, 420

P

Parathyroid gland, 420
Paresis, 62, 112, 404
Passing points, 279
Patrolmen, tests for, 270
Percentile rating, 19
Performance:
 scale:
 Army, 76
 Pintner-Paterson, 73-75
 tests, 72-77
Personal Data Sheet, 358
Personal Inventory, Laird's, 363-
 365
Personal Traits Rating Scale 303
Personality:
 difficulty of measuring, 322-325
 glands in, 413-415
 inventory, 360
 measurement, 321-339
 nature, 321
 schedule, 359
Personnel selection, tests in, 124-
 133, 280-302
Physiological measurement, 25,
 393-454
 character, 387
 emotions, 424-440
 fatigue, 393-402
Picture arrangement test, 76
Picture completion tests, 75, 76
Picture tests, 77-79, 167
Pig iron experiment, 259
Pintner-Cunningham Test, 79

Pintner - Paterson Performance
Scale, 73-75

Pitch, measurement of, 145, 146

Play interests test, 199

Policemen, tests for, 270, 290

Porteus Test, 77-79

Postal tests, 290-302

practicability, 295

reliability, 294

samples, 296-302

validity, 291

Pregnancy tests, 422

Premedical scholarship, 184

Pressey X-O Test, 367-371

Probst Service Rating Scale, 241-250

characteristics, 243-245

development, 241

distribution of ratings, 248

reliability, 249

scoring, 246-247

validity, 249-250

Prodigies, study of, 85-86, 88-90

Product scales, 219-221

Professional aptitude tests, 172-186

Professional school achievement tests (*see* Achievement tests)

Psychasthenia, 359

Psychogalvanic reflex, 388, 433-434

Psychological laboratory, 304

Psychological tests, 14-16

Puzzle block test, 73

Q

Questionnaire, 17, 329-332, 358

Quotient, intelligence (*see* Intelligence quotient)

R

Rating:

percentile, 19

scale, 16, 325-329

interests, 195-196

job efficiency, 241-250

Raw measures, 18

Reaction time, 10, 27, 388, 452-454

Reading Examination, Haggerty, 213-214

Relative measures, 18

Reliability, 19-21

Research Division, Civil Service Commission, 266, 269

Respiration measurements, 402

Revue Philosophique, 29

Rhythm sense, measurement of, 145, 147

S

Scale, rating (*see* Rating scale)

Scatter, in intelligence testing, 112-113

Schizophrenia, 359

Scholarship, low, use of tests in, 120-121

Scholastic Aptitude Test for Medical Schools (*see* Medical aptitude test)

School achievement tests (*see* Achievement tests)

Scientific management, 258

Scopolamine, 438

Seashore musical aptitude test, 144-153

Seguin form board, 73, 74

Selection of employees, measurement in, 259, 280-302

Selective process in school groups, 94

Self-expression, 321

Senile psychosis, 112

Sensitivity, tests of, 446-448

Sensory tests, 25-28, 441-454

Service ratings, 241

Service rating scale, 241-250

Sewing scale, 220

Sex hormone tests, 420-423

Short-answer questions, 40, 210-226

appeal, 226

comprehensiveness, 225

construction, 224

development, 210-211

difficulty, 224

objectivity, 223

reliability, 223

scoring, 225

types, 211-213

Short-answer tests in industry, 271

Shorthand tests, 250

Snellen chart, 447

Social adaptability, 185

Social attributes, measurement of,
334-356
Social Intelligence Test, 15, 197,
200, 332, 334, 335-351
Sociality, 321
Social maturity scale, 351-356
Social quotient, 353
Social relations test, 198
Society of Automotive Engineers,
448, 454
**Sones-Harry High School Achieve-
ment Test,** 222
Special ability tests, 40
Specialized intelligence tests, 269
Special talents, measurement of,
143-161
Sphygmomanometer, 398
Spinal fluid tests, 406-408
Stanford Achievement Test, 222
Stanford Revision of Binet Test
(*see* Binet Test)
Steadiness tests, 448-452
Stenographer-typist tests, 290
Stenquist Mechanical Aptitude
Tests, 164-167, 171
Street-car motormen, tests for,
263-265
Strength tests, 443-446
Substitution test, 75
Superiority, 82-108
definition, 85
development of interest in, 85
examples, 90-91
in school children, 87, 91-94
Terman's study, 88
tests for measuring, 95-108
Syphilitic infection, 62, 404, 407

T

Tabes dorsalis, 404
Talents, measurement of, 143-161
Taylorism, 258
Teaching Aptitude Test, 174
Teaching efficiency, measurement
of, 250-253
Temperament, 321
construction, 273-279, 283
questions:
construction, 277
scale, 330

Temperament (*Cont.*)
selection, 276
Tests (*see* Measurements)
**Thorndike Intelligence Examina-
tion,** 18, 39, 106
Threshold of sensitivity, 446
Thyroid, 62, 409, 418
Time and motion study, 309-311
Time sense, measurement of, 145,
147
Tiredness, 394-395
Trade interest test, 197
Trade tests, 250, 271
Training of personnel, 259, 262
Triangle test, 74
True-false questions, 211
Truth serum, 438-440
Typewriting tests, 250

U

Unreliability of grading, 209-210

V

Validity, 19-21, 278
Vineland maze tests, 77-79
Vineland social maturity scale,
351-356
Vocational Interest Blank, 189
Vocational interest test for college
women, 198
Vocational tests, 10, 40, 124-133,
172-186, 189-196
Vocation-to-vocation rating scale
of interests, 195

W

Wabblemeter, 449-452
Wasserman reaction, 404, 407
Weber's law, 26

X

X-O Test for Emotions, 367-371

Y

Yerkes - Bridges - Hardwick Point
Scale, 33, 204

Index of Names

A

Allen, 454
Allport, F. H., 331, 363
Allport, G. W., 331, 363
Anderson, 311
Ayres, 220-221

B

Benedict, 397
Bernreuter, 360, 362
Bills, 130
Binet, 28-32, 65-72, 90, 110-113, 443
Bingham, 284, 303, 306-309, 318
Blatz, 431
Book, 262
Brace, 441
Broom, 349
Brown, 454
Bryan, 262
Burr, 133
Burt, 196, 311

C

Cady, 362
Cannon, 432-436
Carmichael, 82
Carpenter, 397
Cason, 365-367
Cattell, 9, 10, 12, 27
Chappell, 437
Chave, 388
Clendenning, 414
Clothier, 125
Colloton, 138
Cowdery, 189
Cox, 87, 195
Cunningham, 79

D

Dearborn, 79
Doll, 351-356
Drury, 406

E

Elliot, R. M., 167
Elliott, E. C., 209

F

Farran-Ridge, 406
Fechner, 26
Fere, 434
Fernald, 375
Ferson, 173
Filer, 286, 290
Fracker, 153
Freeman, 138, 374
French, 396
Fryer, 127-128, 187-188, 203-204

G

Galton, 86, 138
Garrett, 349
Garrison, 138
Gilbreth, 258, 310
Goddard, 134
Goldstein, 421

H

Haggerty, 39, 79, 213-214
Hall, 27
Hardwick, 204
Harris, 9
Harry, 222
Harter, 262
Hartshorne, 332, 375, 379, 382, 384
Healy, 134
Helmholtz, 26, 453
Hildreth, 139
Hill, 134
Hollingworth, 153
Holzinger, 139
House, 359
Howard, 153
Hubbard, 125, 131
Humm, 330

Hunt, 197, 252, 335, 347, 349, 351,
396, 406
Hunter, 228, 396

J

Johnson, 77, 153, 445

K

Kellogg, 349
Kelly, 222
Kent, 115-117
Kitson, 195-196
Kohlstedt, 363

L

Laird, 359, 363-365
Landis, 434, 437-438
Larson, 148, 153, 437
Lenoir, 153
Loman, 252

M

MacQuarrie, 169, 349
Marston, 437
Mathews, 359, 361
Mathewson, 125
May, 332, 375, 379, 382, 384
Mazer, 421
McAdory (*see* Siceloff)
McGeoch, 370
McHale, 198
Meier, 154-157
Merriam, 139
Miner, 134
Moss, 174, 185, 197, 228, 252, 270,
335, 396, 406, 449, 455
Münsterberg, 263-265
Murchison, 135
Murdock, 220
Murphy, 117

N

Neyman, 363

O

Omwake, 197, 214, 335
Orr, 378

O'Rourke, 107, 132, 170, 171, 197,
286-291, 349
Otis, 38, 101

P

Paterson, 73, 167, 434, 441
Pearson, 443
Pintner, 73, 79, 88, 92, 122, 134,
136, 137, 349
Poffenberger, 394
Porteus, 77-79
Pressey, L. W., 38
Pressey, S. L., 38, 367-372
Probst, 241-250

R

Rackley, 456-457
Ream, 198
Remmers, 121-122
Roe, 396
Ronning, 214
Roos, 446
Rosanoff, 115-117
Ruch, 222
Rugg, 138

S

Saam, 33
Schwarz, 214
Scott, 125, 262
Scripture, 27
Seashore, 143-153, 154, 447
Shuttleworth, 375
Siceloff, 154, 157-161
Simon, 30, 65
Sones, 222
Stanton, 153
Starch, 209
Stenquist, 162-167
Stoddard, 173
Strang, 349
Strong, 189-194, 203
Swope, 448-449
Sylvester, 32-33
Symonds, 375, 380, 385-386, 429

T

Tarchanoff, 434
Taylor, 258-261
Telford, 241, 270

Terman, 39, 65-72, 88, 90-92, 138,
199, 204, 222

Thorndike, 39, 106, 136, 139, 220,
350

Thurstone, L. L., 107, 173, 359,
362, 388-389

Thurstone, T. G., 359

Tiegs, 209-210

Toops, 119, 197

Tredgold, 60

U

Unger, 133

Upshall, 349

V

Vigouroux, 434

Viteles, 257, 259

Voelker, 375

W

Wadsworth, 330

Weber, 26

Wells, 114

Whipple, 34, 39, 441, 445

Whitely, 370

Willing, 220

Willoughby, 330

Wood, 106, 123, 224, 228-229

Woodworth, 358-359

Woodyard, 161

Wundt, 27

Wyman, 200-203

Y

Yerkes, 35, 39, 204

Z

Zyve, 173

